Journal of Parkinson's Disease 1 (2011) 185–196
DOI 10.3233/JPD-2011-11026
IOS Press

Measurement Properties of the SF-12 Health Survey in Parkinson's Disease

Peter Hagell[a,b,*] and Albert Westergren[a]
[a]*School of Health and Society, Kristianstad University, Kristianstad, Sweden*
[b]*Department of Health Sciences, Lund University, Lund, Sweden*

Abstract. The 12-item Short-Form Health Survey (SF-12) is an abbreviated version of the SF-36, one of the most widely used patient-reported health outcome rating scales. Similar to the SF-36, it yields summary scores of physical and mental health (PCS and MCS, respectively). However, SF-36 derived PCS and MCS scores have not been found valid in neurological disorders such as Parkinson's disease (PD). Here we used modern psychometric methodology (Rasch analysis) to test the SF-12 in PD, and explored the appropriateness of a total SF-12 score representing overall health. SF-12 data from 150 non-demented people with PD (56% men; mean age/PD-duration, 70/5 years) were analyzed regarding Rasch model fit for the PCS, MCS, as well as for the full SF-12. Data showed some signs of misfit to the Rasch model for all three scales (overall item-trait interaction, $P \geq 0.003$; reliability, ≥ 0.85). For example, all scales exhibited signs of dependency between item responses, and the PCS measured with relatively low precision. Model fit (but not measurement precision) was improved following deletion of one PCS and one MCS item (overall item-trait interaction, $P \geq 0.387$; reliability, ≥ 0.82). These observations suggest that the SF-12 can be used as a coarse health survey tool in PD and that a total SF-12 may be useful as a measure of overall health. However, its appropriateness as an outcome measure can be questioned and it is somewhat unclear exactly what the derived scores represent. As such, the SF-12 should probably be considered an assessment tool (or index) rather than a measurement instrument.

Keywords: Health status, outcome assessment, Parkinson disease, psychometrics

INTRODUCTION

The 36-item Short-Form Health Survey (SF-36) [1] is one of the most widely used generic (non-disease specific) patient-reported health outcome rating scales in general as well as in clinical Parkinson's disease (PD) studies [2]. The SF-36 consists of eight scales that are assumed to represent domains of physical and mental health. Based on factor analytic studies on American SF-36 data, it is further assumed that the eight domains can be combined to form two summary measures of physical and mental health (the physical and mental component summary scores, PCS-36 and MCS-36) [3]. These summary measures have been suggested to have advantages over the eight scales by reducing the risk of chance findings and improving the potential to detect clinically significant change [4]. However, studies have challenged the usefulness of the PCS-36 and MCS-36 in several neurological conditions [5–11], including Parkinson's disease (PD) [2, 10, 12]. Specifically, SF-36 domains suggested to represent mental health have been found to have more in common with the PCS-36 than the MCS-36, and vice verse. Furthermore, exploratory [2] and confirmatory [12] factor analyses have suggested that the eight scales conform better to a unidimensional measurement model than to the suggested two-dimensional physical/mental health model.

*Correspondence to: Peter Hagell, P.h.D., School of Health and Society, Kristianstad University, SE-291 88, Kristianstad, Sweden. Tel: +46 44 204056; E-mail: Peter.Hagell@hkr.se.

ISSN 1877-7171/11/$27.50

The 12-item Short-Form Health Survey (SF-12) was developed as an abbreviated alternative to the SF-36 for use in surveys and other situations with constraints on questionnaire length or where more detailed health assessments are not required [13]. The objective of the SF-12 was to reproduce SF-36 derived PCS-36 and MCS-36 scores. This was accomplished by means of regression analyses that identified and weighted the 12 items (representing all eight SF-36 scales) that best reproduced the PCS-36 and MCS-36 [13]. In the original SF-12 scoring algorithm, each item score is weighted by a factor derived from regressing response category scores of each of the SF-12 items on PCS-36 and MCS-36 scores in a general United States sample [13], and each item contributes to both scores in an orthogonal (uncorrelated) manner [13, 14]. The resulting SF-12 has been found to explain around 90% of the variance in PCS-36 and MCS-36 scores [13, 15–19]. However, similarly to the SF-36, investigators have questioned the proposed grouping of items into the PCS-12 and MCS-12 as well as the validity of its orthogonal scoring algorithm, which assumes that physical and mental health are uncorrelated [10, 18–22]. While alternative scoring procedures that do not assume that physical and mental health are uncorrelated have been suggested [20, 22], we are unaware of any assessments of the psychometric performance of the SF-12 if treated as two independent unweighted total scores.

Whereas the SF-36 and SF-12 were developed within the classical test theory framework, modern test theory (particularly the Rasch model) is increasingly considered advantageous in scale development and evaluation [23–27]. Advantages of this approach compared to correlation-based classical test theory approaches include that when responses to a set of items conform to Rasch model expectations, unweighted total scores allow for the construction of interval level measures. If data do not conform to the Rasch model, the analysis provides opportunity to diagnose and understand the problem(s), thereby providing empirically based exploration of remedies such as adjustment of response categories, deletion and regrouping of items. However, no study appears to have used modern psychometric methods such as Rasch analysis to assess the measurement properties of the SF-12 in people with PD.

Here we use Rasch analysis to test the measurement properties of the SF-12 as a measure of physical and mental health among people with PD. In addition, we explore the appropriateness of a single total score measuring overall health.

Table 1
Sample characteristics ($n = 150$)

Gender (men/women)	82 (56.2)/64 (43.8)[a]
Age (years)	70.4 (7.9; 49–85)[b]
Married or cohabitant	106 (72.6)[a]
Disease duration (years)	5.0 (4.9; 0.5–25)[b]
Hoehn & Yahr stage[d]	I (I–II; I–IV)[c]

[a]*n* (%); [b]Mean (standard deviation; min-max); [c]Median (q1–q3; min-max); [d]As assessed for the "on" phase. Range, I–V (*I* = mild unilateral disease; *V* = Confined to bed or wheelchair unless aided) [28].

MATERIALS AND METHODS

Patients

Data were taken from a survey conducted at a PD outpatient clinic at a south Swedish central hospital serving a population of about 170 000 people. The clinic provides multidisciplinary PD care to people representing all stages of the disease according to Hoehn & Yahr [28]. Survey inclusion criteria were people with idiopathic PD without significant cognitive impairment, as assessed according to the Short Test of Mental Status [29]. The study was approved by the local research ethics committee. From an initial sample of 181 eligible people, 159 (88% response rate) consented to participate and 150 (83%) provided useable SF-12 data (Table 1).

The 12-item short form health survey (SF-12)

The SF-12 [13] is an abbreviated version of the SF-36 [1], consisting of 12 items representing the eight health domains covered by the SF-36 [13]. However, due to its brevity (one or two items from each SF-36 domain) it does not produce a profile of the eight SF-36 domains. Instead, it was developed to reproduce the two physical and mental component summary scores (PCS-12 and MCS-12, respectively), assumed to be represented by six items (with two to six response categories) each (Table 2).

Analysis

The PCS-12 and MCS-12 were analyzed individually as two separate six-item physical and mental health scales, respectively. In addition, the appropriateness of a total score representing overall health was examined.

Analyses were conducted according to the Rasch measurement model [24, 25, 30–32]. This model defines, mathematically, what is required from item responses in order to express linear measures.

Table 2
The 12-item short form health survey (SF-12)

Scales	Items		Response categories
	No.	Contents (abridged)	
PCS-12	1	General health	Excellent/Very good/Good/Fair/Poor
	2	Moderate activities	Limited a lot/Limited a little/Not limited at all
	3	Climb several flights of stairs	Limited a lot/Limited a little/Not limited at all
	4	Accomplished less (physical)	Yes/No
	5	Limited in kind of work	Yes/No
	8	Pain - interference	Not at all/A little bit/Moderately/Quite a bit/Extremely
MCS-12	6	Accomplished less (emotional)	Yes/No
	7	Did work less careful	Yes/No
	9	Calm and peaceful	All of the time/Most of the time/A good bit of the time/Some of the time/A little of the time/None of the time
	10	Energy	All of the time/Most of the time/A good bit of the time/Some of the time/A little of the time/None of the time
	11	Downhearted and blue	All of the time/Most of the time/A good bit of the time/Some of the time/A little of the time/None of the time
	12	Social limitations - time	All of the time/Most of the time/Some of the time/A little of the time/None of the time

According to the Rasch model, the probability of a certain item response is a logistic function of the difference between the level of the measured construct represented by the item and that possessed by the person. The model separately locates persons and items 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 location set at zero). The extent to which successful measurement has been achieved is determined by examining the fit between observed data and model requirements. The Rasch model requires that items in a scale reflect a single variable (unidimensionality) and that item responses are independent of each other (local independece). These requirements were tested for the PCS-12 and MCS-12 individually, as well as for a total SF-12 health score. Overall model fit was assessed by examining the mean item and person residual values (i.e., the differences between observed and model expected responses), which should be close to 0 with a standard deviation (SD) close to 1, and the chi-square based item-trait interaction statistic, which should be non-significant. Fit of individual items was analyzed by chi-squared and ANOVA based F-statistics (which should be non-significant [30, 31]) of the residuals across three subgroups (class intervals, CIs) of people defined by their levels of health according to their scores on the respective SF-12 scales. Reliability was estimated by the person separation index (PSI), which is analogous to coefficient alpha [33] and should be $\geq$0.7 for a scales to be able to distinctly differentiate at least two strata of people [34].

Differential item functioning (DIF) is an additional aspect of fit to the Rasch model and an important facet of valid measurement [25, 30, 31, 35]. DIF occurs when items have different meanings and statistical properties across sample subsets, either in a uniform (responses differ uniformly regardless of people's location on the variable) or non-uniform (differences in responses vary across the variable) manner [31, 35]. Analyses of uniform and non-uniform DIF were conducted by testing the hypothesis of no DIF using 2-way ANOVA of the differences in item response functions between genders and age groups (as defined by the median, <71.5 vs. $\geq$71.5 years old) across three CIs.

We then examined whether response categories work as intended, i.e., whether they reflect an increasing amount of the measured variable. If 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) are disordered, these categories do not work as intended. This indicates problems such as too many or overlapping response categories, or may be due to multidimensionality [24, 25, 31].

To assess how well the SF-12 scales accord with the levels of health experienced by the sample the relationships between the locations of persons and items, as determined by Rasch analyses, were examined. If scales are well targeted to the sample, the mean sample location should approximate the mean item location (i.e., zero). A difference of about 0.5 logits (or more) is typically considered meaningful [36, 37]. Targeting is also an important aspect of model fit; when targeting is poor the ability to assess fit is compromised [25, 30, 31]. Similarly, when reliability is low and persons are poorly separated by the scale, the ability to detect misfit lessens [25, 30, 31].

Examination of the relative locations of people and item response category thresholds on the common latent variable also provides a means of assessing the extent to which a scale is successful in mapping out a continuum that represents relevant levels of the measured variable [25, 38]. The person-to-item threshold distributions were therefore examined to determine whether (a) item response thresholds were evenly spread along approximately the same range of the variable as the persons; (b) there were any notable gaps ($\geq$ about 0.5 logits [36, 37]) among the distributions of item response thresholds (indicating compromised measurement ability and larger measurement error); and (c) if item response thresholds tended to cluster together at approximately the same levels on the variable continuum (indicating measurement redundancy).

While assessment of fit to the Rasch model addresses unidimensionality and local independence, residual based fit statistics can be somewhat insensitive in detecting multidimensionality [39, 40]. Smith [39] therefore proposed conducting a principal component analysis (PCA) of the residuals to identify potential subdimensions in the scale, followed by a series of independent *t*-tests to assess whether subsets of items yield different person measures. If violation of unidimensionality is trivial, the number of person locations that differ between two item sets is small. Person location estimates were thus derived from two subsets of items of the respective SF-12 scales that loaded positively and negatively on the first principal component of residuals. The overall proportion of persons with significantly different measures from the two item subsets (or the lower bound of the associated binomial 95% CI) should be <5% to support unidimensionality [39, 41].

Finally, for each scale we explored causes of identified problems and potentials for improvements [31]. First, in case of disordering, response category thresholds were reduced by combining adjacent categories. Second, in case of DIF, this was adjusted for by splitting items into two new items, one for each subgroup involved in the observed DIF [31, 35]. Third, to accommodate items displaying signs of response dependence (violation of local independence) we examined the correlations between item residuals, which should be close to 0 under local independence [25]. Items with high residual correlations were then combined into a "subtest" item, which treats the combined items as one item in the analysis (for example, a subtest created from two dichotomous items would be analyzed as a 3-category polytomous item); this accommodates the response dependence [25]. Response dependence artificially inflates reliability and if reliability notably decreases following the creation of subtests, this signals local dependence and provides a more accurate reliability estimate [25]. Finally, if these remedies did not improve a scale, the effect of removing misfitting items was explored [24, 35].

Analyses were performed using SPSS 14 (SPSS Inc., Chicago, IL) and RUMM2020 (Rumm Laboratory Pty Ltd., Perth). *P*-values are two-tailed and considered significant when <0.05 following Bonferroni adjustment.

RESULTS

Individual SF-12 physical and mental health scales

The PCS-12 showed overall misfit to the Rasch model (Table 3). At the item level, items 4 and 5 showed signs of misfit (Table 4; Fig. 1A, B). Reliability was 0.85 (Table 3). There were no significant DIF between genders or age groups and the response options were working as expected without any threshold disordering. Targeting showed a mean person measure of −0.88 logits, indicating that the sample experienced worse physical health than that conceptualized by the scale. The scale also displayed several gaps along the measured construct (Fig. 2A). The first principal component identified by PCA of residuals explained 30.2% of the variance, and independent *t*-tests showed that 4.8% ($n = 7$; 95% CI, 1.3% to 8.4%) of the person measures derived from the three most positively loading items were significantly different from those derived from the three most negatively loading items (Table 3). This suggests that the two item subsets tap a common variable and therefore supports unidimensionality.

The MCS-12 showed overall fit to the Rasch model (Table 3). However, item 6 showed signs of misfit (Table 4; Fig. 1C). Reliability was 0.85 (Table 3). There were no significant DIF either by gender or age and

Table 3
Overall Rasch model fit statistics, reliability, targeting and unidimensionality of SF-12 derived scores of physical (PCS-12), mental (MCS-12) and overall health (SF-12/SF-10) [a]

	Original scales			Revised scales		
	PCS-12	MCS-12	SF-12	PCS-12[i]	MCS-12[j]	SF-10[k]
Overall Rasch model fit:						
Items						
Fit residual, mean[b]	−0.34	−0.06	0.05	−0.19	0.12	0.20
Fit residual, SD[c]	1.43	1.01	1.18	1.15	1.11	0.64
Persons						
Fit residual, mean[b]	−0.40	−0.29	−0.25	−0.38	−0.28	−0.23
Fit residual, SD[c]	0.83	0.97	1.12	0.85	0.93	1.08
Total item-trait interaction						
Total item chi-square (df)	29.57 (12)	16.32 (12)	38.56 (24)	10.63 (10)	9.75 (10)	19.08 (20)
P-value	0.003[l]	0.177	0.0304[m]	0.387	0.462	0.516
Reliability						
Person separation index[d]	0.85	0.85	0.89	0.82	0.82	0.87
Targeting:						
Person location, mean (SD)[e]	−0.88 (2.20)	0.44 (1.47)	−0.14 (1.44)	0.78 (2.05)	0.55 (1.37)	−0.05 (1.37)
Unidimensionality[f]						
PC1 eigenvalue (% explained variance)[g]	1.81 (30.2)	1.59 (26.6)	2.67 (22.2)	1.80 (36.0)	1.58 (31.6)	2.39 (23.9)
% significantly different person locations (95% CI)[h]	4.8 (1.3, 8.4)	4.9 (1.3, 8.5)	11.3 (7.8, 14.8)	4.8 (1.3, 8.4)	2.8 (−0.8, 6.4)	8.0 (4.5–11.5)

[a] As analysed with the sample divided into three class intervals according to person locations on the measured variable. Data are rounded to two decimals (*P*-values to three decimals, percentages to one decimal); [b] Should be close to 0 [31]; [c] Should be close to 1 [31]; [d] Analogous to Cronbach's alpha [33]; [e] Relative to the mean item logit location (i.e., zero); [f] According to the independent *t*-test protocol [39, 41]; [g] For the first principal component (PC1), as derived from principal component analyses of residuals; [h] Comparison of person location measures derived from items with positive and negative loadings on the first principal component in principal component analyses of residuals [41]; [i] Item 5 omitted; [j] Item 6 omitted and item 11 rescored (from 012345 to 011234); [k] Items 5 and 6 omitted and item 11 rescored (from 012345 to 011234); [l] $P = 0.019$ following Bonferroni correction; [m] $P = 0.364$ following Bonferroni correction.

response category thresholds were ordered. Targeting was better than for the PCS-12 with a mean person measure of 0.44 logits, and some gaps along the measured construct were evident but less prominent than for the PCS (Fig. 2B). The first principal component identified by PCA of residuals explained 26.6% of the variance, and independent t-tests showed that 4.9% ($n = 7$; 95% CI, 1.3% to 8.5%) of the person measures derived from three positively loading items were significantly different from those derived from three items with negative loadings (Table 3). This supports unidimensionality.

Next, we explored potentials for scale improvements. There were two misfitting items in PCS-12 (items 4 and 5). Because these items showed signs of local dependency (negative fit residuals and a residual correlation of 0.476), they were combined into a single subtest item. This improved overall model fit (item mean [SD] fit residual, −0.16 [1.16]; χ^2, 13.25 [df, 10]; $P = 0.210$) and reliability decreased to 0.82, supporting the presence of response dependency. However, the combined subtest item showed misfit (fit residual, −1.1; χ^2, 5.5; F-ratio, 8.48; $P = 0.0003$). We therefore explored the scale further by considering item deletion. Since item 4 was better fitting than item 5 (Table 4; Fig. 1A, B), item 5 was removed. This improved overall model fit (Table 3) and did not introduce any DIF or threshold disordering. Furthermore, the fit of item 4 improved (fit residual, −0.81; χ^2, 4.93; F-ratio, 4.91; $P = 0.043$) and no other misfit was introduced. Targeting improved slightly and unidimensionality remained supported (Table 3).

There was one misfitting item in MCS-12 (item 6). Because the nature of the item statistics suggested response dependency, residual correlations were examined but were all low. We therefore explored the effects of removing item 6 from the scale. Deleting item 6 from the MCS improved overall model fit somewhat (item mean [SD] fit residual, 0.15 [1.02]; χ^2, 8.54 [df, 10]; $P = 0.576$; reliability, 0.82) and did not introduce any

Table 4
Rasch item and fit statistics[a]

No.	Item[b]	Item statistics[c]		Fit statistics		
	Contents (abridged)	Location	SE	Residual[d]	Chi square[e,f]	F ratio[f,g]
PCS-12						
1	General health	1.086	0.159	1.960	5.139	2.264
4	Accomplished less (physical)	0.853	0.241	−0.961	5.681	**6.595**
5	Limited in kind of work	0.250	0.223	−2.020	**14.835**	**24.312**
8	Pain - interference	0.117	0.117	0.164	0.995	0.370
2	Moderate activities	−0.044	0.178	−1.402	2.572	2.508
3	Climb several flights of stairs	−0.925	0.170	0.201	0.346	0.137
MCS-12						
6	Accomplished less (emotional)	0.838	0.200	−1.569	**12.041**	**11.066**
10	Energy	0.707	0.094	0.338	0.214	0.071
7	Did work less careful	0.046	0.195	1.420	0.400	0.111
9	Calm and peaceful	−0.158	0.092	−0.443	2.281	1.566
11	Downhearted and blue	−0.685	0.093	−0.410	1.158	0.565
12	Social limitations - time	−0.747	0.105	0.335	0.227	0.094
SF-12						
1	General health	1.516	0.142	0.357	2.841	1.310
4	Accomplished less (physical)	1.302	0.223	−0.860	3.837	2.470
5	Limited in kind of work	0.794	0.205	−1.971	5.767	5.174
2	Moderate activities	0.476	0.156	−0.043	0.297	0.272
6	Accomplished less (emotional)	0.222	0.194	−2.112	**11.913**	**10.500**
10	Energy	0.113	0.094	0.160	1.509	0.785
3	Climb several flights of stairs	−0.195	0.147	1.070	1.069	0.259
8	Pain - interference	−0.412	0.101	0.448	2.395	1.158
7	Did work less careful	−0.503	0.192	0.752	3.340	1.550
9	Calm and peaceful	−0.734	0.092	1.231	2.503	1.239
11	Downhearted and blue	−1.214	0.093	1.622	1.960	0.790
12	Social limitations - time	−1.364	0.107	−0.049	1.127	0.565

[a]Performed with the sample divided into three class intervals according to person locations on the measured variable; [b]Listed in location order, from better to poorer health; [c]Expressed in linear log-odds units (logits), with mean item location set at 0 for each scale; [d]Log residuals summarise the deviation of observed from expected responses. Deviation from the recommended [31] range of −2.5 to +2.5, indicating item misfit, are bold; [e]Chi square values summarise the deviation of observed from expected responses across the three class intervals of the sample; [f]Bonferroni corrected statistically significant deviations across class intervals, indicating item misfit, are bold; [g]One-way ANOVAs of deviations from model expectation across the three class intervals of people; SE = standard error.

item misfit or DIF but response category thresholds for item 11 showed some disordering (Fig. 3A). Item 11 was then rescored (from 012345 to 011234; Fig. 3B). This did not yield any unequivocal further change in overall model fit (Table 3), and did not introduce any DIF or threshold disordering. Targeting was virtually unchanged and unidimensionality remained supported (Table 3).

Fit of the total SF-12 as a measure of overall health

When treating all 12 items as an overall measure of health the SF-12 showed overall fit to the Rasch model (Table 3), no DIF, but signs of misfit for item 6 (Table 4). Reliability was 0.89 (Table 3). Item 11 also displayed disordered thresholds (same pattern as in Fig. 3A). Targeting was good (Table 3) with few gaps along the measured construct, apart for the upper end of the scale where those in best health were measured with relatively low precision (Fig. 2C). The first principal component identified by PCA of residuals explained 22.2% of the variance. Interestingly, this analysis showed that all PCS-12 items loaded positively on the first principal component, whereas all MCS-12 items had a negative loading (Table 5). Independent t-test comparisons of the person measures from these two item sets showed that 11.3% (95% CI, 7.8% to 14.8%) of the people had significantly different measures depending on whether they were measured with PCS-12 or MCS-12 items.

To explore potentials for improvement of the scale, item 11 was rescored (from 012345 to 011234) and the scale was reanalyzed. This improved overall model fit

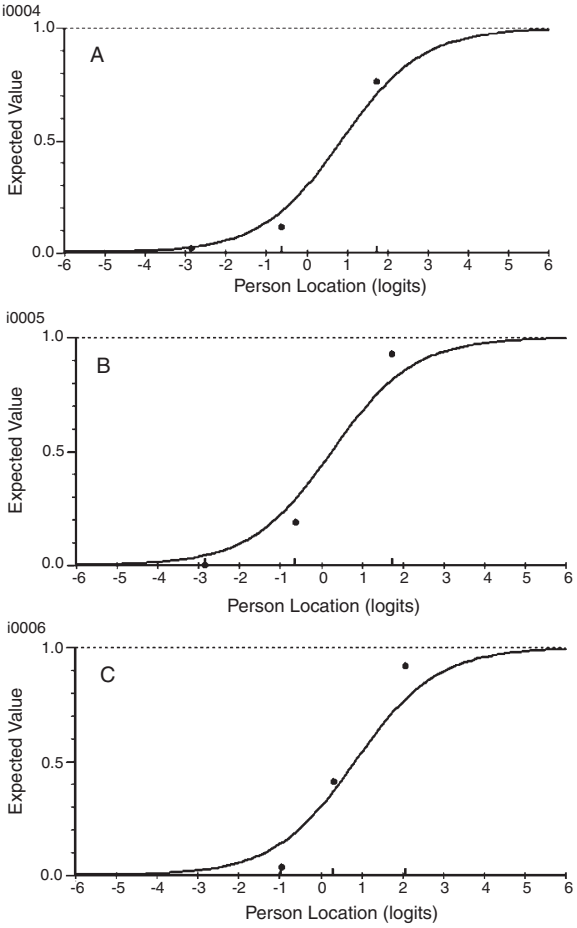

Fig. 1. Item characteristic curves (ICCs) of SF-12 items. ICCs (grey curves) represent the expected item responses (y-axis) at various levels of the measured construct (x-axis). Black dots represent the observed responses in the sample as divided into three class intervals according to their locations on the measured construct, indicated by the marks on the x-axis. A. item 4 ("accomplished less due to physical health") of the PCS-12; B. item 5 ("limited in kind of work") of the PCS-12; C. item 6 ("accomplished less due to emotional health") of the MCS-12. All three panels illustrate instances of over discrimination as the empirical observations are steeper than expected. This suggests that items may be redundant and do not contribute unique information in addition to that already provided by other item(s). SF-12, The 12-item Short-Form Health Survey; PCS, Physical Component Summary scale; MCS, Mental Component Summary scale.

somewhat (item mean [SD] fit residual, 0.03 [1.16]; χ^2, 34.54 [df, 24]; $P = 0.075$; reliability, 0.89) and did not introduce any DIF, whereas item 6 still displayed misfit (fit residual, −2.06; χ^2, 10.11; F-ratio, 8.98; $P = 0.0024$). Examination of residual correlations for this item displayed signs of response dependency with item 7 (correlation, 0.32). These items were therefore combined into a subtest, which improved overall model fit (item mean [SD] fit residual, 0.14 [0.96]; χ^2, 23.95 [df, 22]; $P = 0.350$) and reliability decreased to 0.88. However, there were still signs of multidimensionality as *t*-test comparisons of the person measures from the two item sets showed that 10.3% (95% CI, 7.2% to 17.8%) of the people had significantly different measures depending on the item subset used. We therefore explored the scale further by considering item deletion.

Removal of item 6 from the scale improved overall model fit slightly (item mean [SD] fit residual, 0.11 [0.98]; χ^2, 31.28 [df, 22]; $P = 0.091$; reliability, 0.88), did not introduce any DIF, but caused item 5 to display misfit (fit residual, −2.0; χ^2, 6.90; F-ratio, 6.48; $P = 0.022$). Following removal also of item 5, overall fit improved (Table 3) with no DIF or item misfit. Targeting was somewhat improved compared to the original SF-12 (Table 3) but, as expected, there were more gaps along the measured construct (Fig. 2D). Unidimensionality was now statistically supported according to the independent *t*-test comparisons of person measures from items with positive (items 1–4 and 8) and negative (items 7 and 9–12) loadings on the first principal component (Table 3).

DISCUSSION

This is the first study to use Rasch analysis to assess the measurement properties of the SF-12 among people with PD. Our data provide some evidence that raw SF-12 item scores can be used to assess aspects of physical, mental and overall health in PD. However, we also identified aspects of the instrument that could be improved, and it is still somewhat unclear exactly what the derived measures represent.

There are several findings in this study that support the measurement properties of the SF-12. Item misfit was relatively benign for both the PCS-12 and the MCS-12, and reliability was acceptable. Items also appear to work the same way among men and women and older and younger people with PD, which provides the foundation for fair comparisons between these groups of people. We did, however, find indications that the six response categories for item 11 were not working as assumed. Specifically, respondents appear to experience problems distinguishing between feeling downhearted and blue "a little of the time" and "some of the time". As indicated by experiences from other health outcome scales [25, 35] and by our findings from treating these two response categories as one, reducing the number of response alternatives may render their distinction clearer and result in improved

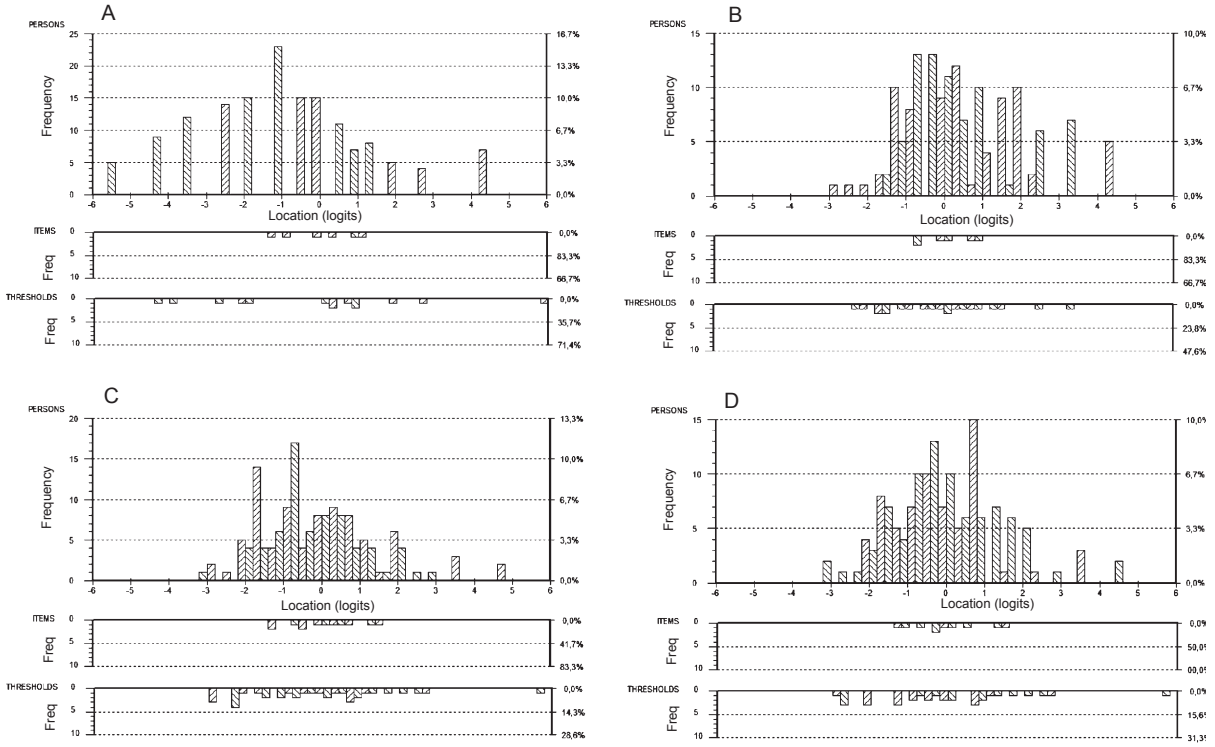

Fig. 2. Distribution of the locations of people (upper panels), SF-12 items and response category thresholds (first and second lower panels, respectively) on the common logit metric (x-axis; positive values = better health). A, PCS-12; B, MCS-12; C, SF-12 total score; D, SF-12 total score following deletion of items 5 and 6. SF-12, The 12-item Short-Form Health Survey; PCS, Physical Component Summary scale; MCS, Mental Component Summary scale.

measurement. However, the disordering was relatively minor and other items using the same response scale did not display disordered thresholds. Further studies in larger samples are therefore needed before any firm conclusions can be drawn.

We found evidence in support for the unidimensionality of the PCS-12 and MCS-12. Neither item fit statistics nor the PCA based independent *t*-test protocol indicated any obvious signs of multidimensionality. This is encouraging since unidimensionality is a requirement for valid summation of item scores into total scores and violation thereof renders the meaning of total scores ambiguous [39]. We did, however, find evidence for some local dependency, primarily involving items 5 and 6. Although this type of violation to the measurement model may be thought of as less problematic than multidimensionality, research has shown that response dependence can influence person measures derived from a scale and mask differences when measuring change [42, 43]. While replication studies are needed, our observations suggest that this should be taken into account and that the questionnaire potentially may benefit from removing items 5 and 6, at least when used among people with PD. It is recognized that this observation may not translate into other populations and it might be considered impractical to have different versions of a scale for different groups of people. However, as long as it can be shown that items represent the same variable it is, in principal not a problem to use different items in different situations [44]. It must also be emphasized that any data manipulation (e.g., collapsing categories or deleting items) are exploratory post hoc exercises and it is unknown how the questionnaire actually had worked with fewer items or if an item had fewer response categories or less items. The main interpretation should therefore be limited to results from the original questionnaire.

In accordance with observations on the relationships among the eight SF-36 scales in PD [2] and the scoring of the RAND version of SF-36 [45], we also explored whether the SF-12 could be used to derive a single measure of overall health. The results were very similar to those from the analyses of the PCS-12 and MCS-12 in that it generally behaved well, but item 11 displayed

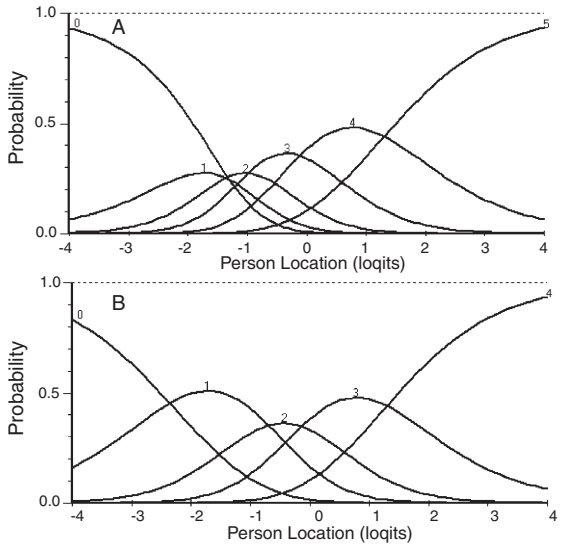

Fig. 3. Response category probability curves for item 11 ("downhearted and blue") of the MCS-12. Location on the measured construct is indicated on the x-axis (with threshold locations centered at zero; positive values = better health) and the y-axis represents the probability of affirming response categories 0 ("all of the time"), 1 ("most of the time"), 2 ("a good bit of the time"), 3 ("some of the time"), 4 ("a little of the time"), and 5 ("none of the time"). Category probability curves show the probability of observing each category relative to the location on the measured construct (x-axis). A. original response categories displayed disordered thresholds between categories 0-to-1 and 1-to-2. B. following rescoring (from 012345) by collapsing categories 1 and 2 (011234), category thresholds were ordered as expected.

Table 5
Principal component analysis of residuals of the SF-12[a]

No.	Item[b]	PC1 loadings
	Contents (abridged)	
9	Calm and peaceful	0.697
11	Downhearted and blue	0.642
7	Did work less careful	0.141
12	Social limitations - time	0.119
10	Energy	0.096
6	Accomplished less (emotional)	0.050
1	General health	−0.096
8	Pain - interference	−0.375
4	Accomplished less (physical)	−0.536
3	Climbing several flights of stairs	−0.616
5	Limited in kind of work	−0.660
2	Moderate activities	−0.687

[a]Performed with the sample divided into three class intervals according to person locations on the measured variable; [b] Ordered from the largest positive to the largest negative loading; PC1, the first principal component.

some response category disordering and items 5 and 6 displayed signs of local dependency. However, the independent *t*-test protocol suggested multidimensionality, which disappeared following removal of items 5 and 6. It is interesting to note that PCS-12 and MCS-12 items loaded at opposite ends on the first principal component (this pattern was preserved in the PCA after removal of items 5 and 6; data not shown). This supports the view that PCS-12 and MCS-12 items represent different aspects of a common variable.

In all analyses, there was support that items and response category thresholds mapped out a continuum along the latent variable, ranging approximately 10 (PCS-12), 5 (MCS-12) and 9 (SF-12 total score) logits. However, there were typically gaps along these ranges, particularly for the PCS-12. This means that measurement precision is relatively low and measurement errors are relatively large. As a consequence, the scale will have less ability to accurately locate people, distinguishing between subgroups and detect changes in health status along intervals of the continuum not represented by response category thresholds [25, 46]. The extent to which this is a problem will depend on the study objectives at hand; if used as a survey tool in order to get a rough idea about peoples' health it may be acceptable, whereas it may not be if the goal is to detect smaller but clinically important differences or changes in a clinical trial.

A final aspect to consider is that of the internal structure of the scales. That is, are items located in a clinically reasonable hierarchy from less to more of the target variable, and what variable do they represent? As PD is a chronically progressive disorder, it may be reasonable to consider the SF-12 item locations as progressive hierarchies from positive (better health) to negative locations. To ease interpretation, items in Table 4 are listed in this fashion. Thus, according to the PCS-12, physical health impairment in PD would first manifest as compromised general health perception, then inability to accomplish as much as desired due to physical health problems, limitations in the kind of work/activities performed, pain that interferes with normal work, limitations in moderate activities and, finally experiencing problems climbing stairs. In general, this pattern appears clinically reasonable. For example, it is not uncommon that people with PD who experience walking difficulties still find it relatively easy to climb stairs [47]. Similarly, the hierarchical item structure of the MCS-12 appears generally reasonable. Recent studies have, for example, suggested that fatigue (lack of energy) is an early, frequent and distressing feature of PD [48]. Finally, when analyzed as a single SF-12 total score the hierarchical ordering of items suggest that mental health aspects generally appear to be associated with worse health status relative to physical health aspects.

When items fit the Rasch model the differences in locations (not the actual locations, which always are relative to the arbitrary zero-point on the logit scale) are expected to be invariant. Examination of the item locations in Table 4 also reveals that, with only one exception (item 8) the relative item hierarchies within the PCS-12 and MCS-12 are preserved when all 12 items are analyzed as a total SF-12 score. The differences in locations between successive item pairs in the PCS-12 and MCS-12 hierarchies were also preserved (within allowable error ranges) when all 12 items were analyzed as a total SF-12 score, except for differences involving items 5, 6 and 8. This adds to the concerns regarding the appropriateness of items 5 and 6.

A more intricate question for consideration is what variable(s) the SF-12 represents; that is, what variable(s) manifest it-/themselves with the health problems expressed by the 12 items? Arguable, whereas the MCS-12 items appear to be reasonable manifestations of emotional/affective problems, it appears more difficult to appreciate a single definable latent variable that manifests itself in terms of activity limitations, mobility problems, pain and general health perception. Such concerns do not diminish when considering all 12 items together. One interpretation is that the SF-12 primarily is to be regarded an assessment tool (or index) rather than an instrument for measurement. That is, in contrast to what is assumed according to classic and modern psychometric theory, item responses are not manifestations of (caused by) varying levels in the latent variable they represent but the other way around [49–52]. This cannot be determined through Rasch analysis or other common psychometric methods [50, 51].

Although the sample here is within the general limits of what is required for stable item estimates [53], a limitation of the present study is the relatively small sample size. The importance of sample size in Rasch analysis relates to targeting and sample size requirements increase as targeting deteriorates [53]. It should therefore be noticed that (with one exception in three analyses) all item thresholds in all analyses were covered by the sample (Fig. 2). This increases the confidence in the estimated item parameters. However, additional studies in larger samples are required for firmer conclusions. Furthermore, the PCA/*t*-test protocol for testing unidimensionality is sensitive to the number of thresholds in the scale and should therefore be interpreted with some caution, particularly for the PCS-12 and MCS-12.

CONCLUSIONS

This is the first study to evaluate the measurement properties of the SF-12 using Rasch analysis. Our observations suggest that the SF-12 can be useful as a coarse health survey tool in PD and that a total SF-12 score may be useful as a measure of overall health. However, evidence suggests that some of its items may compromise the scale in this population and it is somewhat unclear what variable(s) the SF-12 represents. As such, it is likely to primarily be useful as an assessment tool or index rather than a measurement instrument. Further studies in larger samples are warranted to investigate these issues in more detail. Based on the results presented here, it can be concluded that the SF-12 may be useful as an assessment/index tool in PD and, as such, it appears more suited for audit situations than for clinical trials. However, it still remains to be determined whether SF-12 data from people with PD are comparable to those from people with other disorders. Before such evidence have been produced, caution is advised in any such use of the questionnaire.

COMPETING INTERESTS

None.

AUTHORS' CONTRIBUTIONS

PH conceived and designed the study, carried out the analyses and drafted the manuscript. AW participated in the conception and design of the study, revision of the manuscript and interpretation of data. Both authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors wish to thank all participating patients for their cooperation and Susanne Lindskov for assistance with data collection. The study was conducted within the Basal Ganglia Disorders Linnaeus Consortium (BAGADILICO) at Lund University, Sweden, and within the Clinical Assessment Research and Education (PRO-CARE) group, Kristianstad University. The study was supported by the Swedish Research Council, the Swedish Parkinson Foundation, the Skane County Council Research and Development Foundation, the Swedish Parkinson Academy, and the Faculty of Medicine at Lund University.

REFERENCES

[1] Ware JE Jr & Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*, **30**, 473-483.

[2] Hagell P, Törnqvist AL & Hobart J (2008) Testing the SF-36 in Parkinson's disease: Implications for reporting rating scale data. *J Neurol*, **255**, 246-254.

[3] Ware JE Jr, Kosinski M & Gandek B (1993, 2000) *SF-36 health survey: Manual and interpretation guide*, QualityMetric Incorporated, Lincoln, RI.

[4] Ware JE Jr & Kosinski M (2001) *SF-36 physical and mental health summary scales: A manual for users of version 1. Second edition*, QualityMetric Incorporated, Lincoln, RI.

[5] Baron R, Elashaal A, Germon T & Hobart J (2006) Measuring outcomes in cervical spine surgery: Think twice before using the SF-36. *Spine*, **31**, 2575-2584.

[6] Cano SJ, Thompson AJ, Bhatia K, Fitzpatrick R, Warner TT & Hobart JC (2007) Evidence-based guidelines for using the Short Form 36 in cervical dystonia. *Mov Disord*, **22**, 122-126.

[7] Dallmeijer AJ, Dekker J, Knol DL, Kalmijn S, Schepers VP, de Groot V, Lindeman E, Beelen A & Lankhorst GJ (2006) Dimensional structure of the SF-36 in neurological patients. *J Clin Epidemiol*, **59**, 541-543.

[8] Hobart J, Freeman J, Lamping D, Fitzpatrick R & Thompson A (2001) The SF-36 in multiple sclerosis: Why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry*, **71**, 363-370.

[9] Hobart JC, Williams LS, Moran K & Thompson AJ (2002) Quality of life measurement after stroke: Uses and abuses of the SF-36. *Stroke*, **33**, 1348-1356.

[10] Jakobsson U, Westergren A, Lindskov S & Hagell P (2011) Construct validity of the SF-12 in three different samples. *J Eval Clin Pract*.

[11] Jenkinson C, Hobart J, Chandola T, Fitzpatrick R, Peto V & Swash M (2002) Use of the short form health survey (SF-36) in patients with amyotrophic lateral sclerosis: Tests of data quality, score reliability, response rate and scaling assumptions. *J Neurol*, **249**, 178-183.

[12] Banks P & Martin CR (2009) The factor structure of the SF-36 in Parkinson's disease. *J Eval Clin Pract*, **15**, 460-463.

[13] Ware J, Jr, Kosinski M & Keller SD (1996) A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Med Care*, **34**, 220-233.

[14] Ware JJr, Kosinski M & Keller SD (1995) *SF-12: How to score the SF-12 physical and mental health summary scales. 2nd ed*, The Health Institute, New England Medical Center, Boston, MA.

[15] Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner JB, Brazier JE, Bullinger M, Kaasa S, Leplege A, Prieto L & Sullivan M (1998) Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: Results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol*, **51**, 1171-1178.

[16] Jenkinson C & Layte R (1997) Development and testing of the UK SF-12 (short form health survey). *J Health Serv Res Policy*, **2**, 14-18.

[17] Nortvedt MW, Riise T, Myhr KM & Nyland HI (2000) Performance of the SF-36, SF-12, and RAND-36 summary scales in a multiple sclerosis population. *Med Care*, **38**, 1022-1028.

[18] Wilson D, Tucker G & Chittleborough C (2002) Rethinking and rescoring the SF-12. *Soz Praventivmed*, **47**, 172-177.

[19] Pickard AS, Johnson JA, Penn A, Lau F & Noseworthy T (1999) Replicability of SF-36 summary scores by the SF-12 in stroke patients. *Stroke*, **30**, 1213-1217.

[20] Farivar SS, Cunningham WE & Hays RD (2007) Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V.I. *Health Qual Life Outcome*, **5**, 54.

[21] Fleishman JA, Selim AJ & Kazis LE (2010) Deriving SF-12v2 physical and mental health summary scores: A comparison of different scoring algorithms. *Qual Life Res*, **19**, 231-241.

[22] Hays RD, Sherbourne CD & Mazel RM (1993) The RAND 36-Item Health Survey 1.0. *Health Econ*, **2**, 217-227.

[23] Hagell P & Nygren C (2007) The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: Implications for evidence based medicine. *J Neurol Neurosurg Psychiatry*, **78**, 1191-1198.

[24] Hagquist C (2001) Evaluating composite health measures using Rasch modelling: An illustrative example. *Soz Praventivmed*, **46**, 369-378.

[25] Hobart J & Cano S (2009) Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technol Assess*, **13**(iii, ix-x), 1-177.

[26] Tennant A, McKenna SP & Hagell P (2004) Application of Rasch analysis in the development and application of quality of life instruments. *Value Health*, **7**(Suppl 1), S22-S26.

[27] Wilson M (2005) *Constructing measures: An item response modelling approach*, Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

[28] Hoehn MM & Yahr MD (1967) Parkinsonism: Onset, progression and mortality. *Neurology*, **17**, 427-442.

[29] Kokmen E, Smith GE, Petersen RC, Tangalos E & Ivnik RC (1991) The short test of mental status. Correlations with standardized psychometric testing. *Arch Neurol*, **48**, 725-728.

[30] Andrich D (1988) *Rasch models for measurement*, Sage Publications, Inc, Beverly Hills.

[31] Andrich D, Sheridan B & Luo G (2004-2005) *Interpreting RUMM*, RUMM Laboratory Pty Ltd., Perth.

[32] Rasch G (1960) *Probabilistic models for some intelligence and attainment tests*, Danmarks Paedagogiske Institut Copenhagen.

[33] Andrich D (1982) An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, **9**, 95-104.

[34] Smith EV Jr (2001) Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *J Appl Meas*, **2**, 281-311.

[35] Hagquist C & Andrich D (2004) Is the Sense of Coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. *Personality and individual differences*, **36**, 955-968.

[36] Hudgens S, Dineen K, Webster K, Lai JS & Cella D (2004) Assessing statistically and clinically meaningful construct deficiency/saturation: Recommended criteria for content coverage and item writing. *Rasch Measurement Transactions*, **17**, 954-955.

[37] Lai JS & Eton DT (2002) Clinically meaningful gaps. *Rasch Measurement Transaction*, **15**, 850.

[38] Stone MH, Wright BD & Stenner AJ (1999) Mapping variables. *J Outcome Meas*, **3**, 308-322.

[39] Smith EV Jr (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas*, **3**, 205-231.

[40] Smith RM (1996) A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, **3**, 25-40.
[41] Tennant A & Pallant J (2006) Unidimensionality matters. *Rasch Measurement Transactions*, **20**, 1048-1051.
[42] Marais I (2009) Response dependence and the measurement of change. *J Appl Meas*, **10**, 17-29.
[43] Marais I & Andrich D (2008) Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *J Appl Meas*, **9**, 105-124.
[44] Bode RK, Lai JS, Cella D & Heinemann AW (2003) Issues in the development of an item bank. *Arch Phys Med Rehabil*, **84**, S52-S60.
[45] Hays RD, Prince-Embury S & Chen HY (1998) *RAND-36 health status inventory*, The Psychological Corporation, San Antonio.
[46] Wright BD & Masters GN (1982) *Rating scale analysis*, MESA Press, Chicago.
[47] Sawle G (1999) *Movement disorders in clinical practice*, Isis Medical Media Ltd., Oxford.
[48] Friedman JH, Brown RG, Comella C, Garber CE, Krupp LB, Lou JS, Marsh L, Nail L, Shulman L & Taylor CB (2007) Fatigue in Parkinson's disease: A review. *Mov Disord*, **22**, 297-308.
[49] Bollen K & Lennox R (1991) Conventional wisdom on measurement: A structural equation perspective. *Psychol Bull*, **110**, 305-314.
[50] Fayers PM & Hand DJ (2002) Causal variables, indicator variables and measurement scales: An example from quality of life. *J R Statist Soc A*, **165**, 233-266.
[51] Stenner AJ, Burdick DS & Stone MH (2008) Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Meas Trans*, **22**, 1152-1153.
[52] Stenner AJ, Stone MH & Burdick DS (2009) Indexing vs. measuring. *Rasch Meas Trans*, **22**, 1176-1177.
[53] Linacre JM (1994) Sample size and item calibration (or person measure) stability. *Rasch Measurement Transaction*, **7**, 328.