The 39-item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence-based medicine

Peter Hagell and Carita Nygren

J. Neurol. Neurosurg. Psychiatry published online 18 Apr 2007;
doi:10.1136/jnnp.2006.111161

JNNP Online First, published on April 18, 2007 as 10.1136/jnnp.2006.111161

The 39-item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence-based medicine

Peter Hagell[1,2,3], Carita Nygren[1]

[1] Department of Health Sciences, Lund University, Lund, Sweden
[2] Department of Neurology, University Hospital, Lund, Sweden
[3] The Vårdal Institute, the Swedish Institute for Health Science, Lund University, Lund, Sweden

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

Key words:
Clinical trials, Parkinson's disease, psychometrics, Rasch analysis, rating scales

ABSTRACT

Background: The PDQ-39 is the most widely used patient-reported rating scale in Parkinson's disease. However, several fundamental measurement assumptions necessary for confident use and interpretation of the eight PDQ-39 scales have not been fully addressed.

Methods: Postal survey PDQ-39 data from 202 people with Parkinson's disease (54% men; mean age 70) were analyzed regarding psychometric properties using traditional and Rasch measurement methods.

Results: Data quality was good (mean missing item responses, 2%) and there was general support for the legitimacy of summing items within scales without weighting or standardization. Score reliabilities were adequate (Cronbach's alpha: 0.72-0.95; test-retest: 0.76-0.93). The validity of the current grouping of items into scales was not supported by scaling success rates (mean, 56.2%), or factor and Rasch analyses. All scales represented more health problems than that experienced by the sample (mean floor effect, 15%), and showed compromised score precision towards the less severe end.

Conclusions: Results provide general support for the acceptability and reliability of the PDQ-39. However, they also demonstrate limitations that have implications for the use of the PDQ-39 in clinical research. The grouping of items into scales appears overly complex and the meaning of scale scores is unclear, which hampers their interpretation. Suboptimal targeting limits measurement precision and, therefore, probably also responsiveness. These observations have implications for the role of the PDQ-39 in clinical trials and evidence-based medicine. PDQ-39 derived endpoints should be interpreted and selected cautiously, particularly regarding small but clinically important effects among people with less severe problems.

The past decade has seen two major developments in clinical Parkinson's disease (PD) research: an increasing focus on evidence-based medicine and a growing emphasis on the importance of patient-reported outcomes.[1 2] It is therefore reasonable to expect the effectiveness of therapy to increasingly be judged on the basis of patient completed rating scales. A prerequisite for valid interpretation of clinical findings and, hence, evidence-based medicine is that rating scales can be interpreted with confidence.[3-6] The need for high quality patient reported rating scales in PD and the fundamental role of evidence-based measurement in clinical research is thus apparent.

The 39-item PD questionnaire (PDQ-39)[7] is the most widely used disease-specific patient completed rating scale in PD.[8] However, several important measurement properties of the PDQ-39 have not been fully addressed. For example, basic requirements (scaling assumptions) that determine the legitimacy of summing PDQ-39 item scores without weighting or standardization have not been examined, and studies addressing the validity of grouping items into its eight scales (dimensionality) have shown inconclusive or discouraging results.[9-12] This poses limitations on the possibility to interpret study outcomes since it may be unclear what scores represent.[4] There have also been indications that the PDQ-39 may not target respondents adequately, which could affect its ability to detect clinically relevant changes.[10] Re-evaluation of the PDQ-39 therefore appears warranted to help inform its use and role in clinical trials and evidence-based medicine.

With this in mind, we assessed the scaling assumptions, reliability, dimensionality and targeting of the eight PDQ-39 scales. Whereas the PDQ-39 was developed within the traditional test theory framework, modern test theory (particularly the Rasch model) is increasingly considered advantageous in scale development and evaluation.[3 13-16] The PDQ-39 was therefore analyzed using both traditional and Rasch measurement methods.

METHODS

Patients and data collection

A total of 451 people with clinically diagnosed PD[17] seen at a South Swedish university hospital during one year was considered for inclusion. Participants in other recent or ongoing questionnaire studies (n=164) were excluded, as well as those deceased or in terminal care (n=30). The remaining 257 people were sent a questionnaire booklet including the Swedish version of the PDQ-39.[10 18 19] Two weeks later a second copy was administered, including a question asking if their health had changed (according to a 5-grade scale, "much better" – "better" – "unchanged" – "worse" – "much worse") since the first mailing. Reminders were sent to non-responders one week after each mailing. Survey response was interpreted as consent to participate. The study was approved by the local research ethics committee.

The first mailing had a response rate of 81% (n=209). Those indicating that they had not answered the survey themselves (n=7) were excluded from further analyses, leaving 202 eligible cases (Table 1). All but seven patients received levodopa with or without adjunct anti-parkinsonian drugs, 18 had undergone neurosurgical interventions for their PD, three were only on PD drugs other than levodopa, and four were not yet on any medical therapy. Of 173 responses to the second mailing (response rate: 67%), five had not responded themselves and 31 reported change in their health status since the first occasion.

Table 1. Sample characteristics (n=202) [a]

Gender (men / women)	108 (53.5) / 94 (46.5) [b]
Age (years)	69.8 (10.0; 34-90) [c]
Retired	143 (70.8) [b]
Married or cohabitant	144 (71.2) [b]
Living in own home	179 (88.6) [b]
Disease duration (years)	8.7 (6.6; 0.5-28) [c]
Hoehn & Yahr stage of PD [e]	III (II-IV; I-V) [d]
Perceived disease severity [f]	2 (2-2; 1-3) [d]
Motor fluctuations [g]	137 (67.8) [b]
Dyskinesias [g]	99 (49) [b]

[a] At time 1 (patients reporting that they had answered the questionnaires themselves).
[b] n (%).
[c] Mean (standard deviation; min-max).
[d] Median (q1-q3; min-max).
[e] As assessed for the "off" phase. Range, I-V (I = mild unilateral disease; V = Confined to bed or wheelchair unless aided).[20]
[f] Self-rated as mild (=1), moderate (=2), or severe (=3).
[g] Self-reported as present or absent.
PD, Parkinson's disease.

The PDQ-39

The PDQ-39 is a PD specific health status questionnaire comprising 39 items proposed to represent eight domains (scales) consisting of three to ten items each (Table 2).[7] Respondents are requested to affirm one of five response categories according to how often (from never to always), due to their PD, they have experienced the problem defined by each item during the past month. The eight PDQ-39 scale scores are generated by Likert's[21] method of summated ratings, i.e., item responses are summed without weighting or standardization. Scores are then transformed to a common range of 0 to 100 (100 = maximum level of problems).

Analyses

Data quality, scaling assumptions and reliability

First, data quality (the percent missing data) was examined. We then examined the scaling assumptions, i.e., the legitimacy of adding up items to generate scores without weighting or standardization.[21] Briefly, these require that within each scale, item scores should have roughly similar means and variances, and that the corrected item-total correlation (i.e., the correlation between each item and the total score of the remaining items in that scale) should exceed 0.4.[22] Internal consistency reliability was assessed by Cronbach's alpha.[23] Test-retest reliability between data from the first and second mailings among respondents who reported stable health (n=137) was assessed by the intra-class correlation coefficient. Reliability estimates should not be below 0.7 and preferably ≥0.8.[24 25]

Dimensionality

Four approaches were used to test whether the proposed grouping of items into eight scales was empirically supported. First, scaling success rates were examined. Scaling success is supported when items correlate significantly stronger with the total score of the other items in their proposed scale (corrected item-total correlations) than with other scales, as determined by 95% confidence intervals.[22] Scaling failure is implied if an item correlates stronger with a scale other than its proposed one.

Items were then subjected to exploratory factor analysis with varimax rotation. Results were first interpreted by the criterion originally used to define the eight PDQ-39 scales,[7] i.e., by retaining factors (scales) with eigenvalues exceeding 1. However, because this criterion tends to overestimate the number of factors, parallel analysis was also used.[26] One thousand parallel sets of random PDQ-39 data were thus generated and factor analyzed, and each

consecutive empirical factor with an eigenvalue exceeding the 95th percentile of random data eigenvalues was considered a useful factor.[27]

Third, the extent by which observed data fitted the hypothesized items-to-scales structure was explored using confirmatory factor analysis. This technique is generally recommended over exploratory factor analysis when there is an *a priori* hypothesis regarding dimensionality, since it allows for testing whether empirical data fit an assumed structure.[28]

Finally, each of the eight proposed PDQ-39 scales were individually examined by means of the Rasch measurement model.[29] According to this 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). A fundamental Rasch model assumption is that all items in a scale work in harmony to define a common unidimensional construct. This assumption was tested for each of the eight PDQ-39 scales through assessment of overall scale and item level model fit by examining the accordance between expected and observed responses.[30] Differential item functioning (DIF) is an additional aspect of fit to the Rasch model and an important facet of valid measurement.[13 30] DIF occurs when items have different meanings and statistical properties across sample subsets. Presence of DIF challenges the validity of comparing data across such subgroups, and threats unidimensionality. DIF was assessed by comparing item response functions between genders and age groups (as defined by the median, <72 vs. ≥72 years old) across various locations on the measured constructs.[13 30]

Targeting

To assess how well the eight PDQ-39 scales[7] accord with the levels of health problems experienced by the sample we first examined the amounts of floor and ceiling effects. That is, the percentage of respondents obtaining the lowest and highest possible scores, respectively, which should not exceed 15%.[31] In addition, 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).

Analyses were performed using SPSS 12 (SPSS Inc., Chicago, IL), ScoreRel CI,[32] AMOS 5 (SmallWaters Corp., Chicago, IL) and RUMM2020 (Rumm Laboratory Pty Ltd., Perth). P-values are two-tailed and considered significant when <0.05.

RESULTS

Data Quality, Scaling Assumptions and Reliability

Data quality was good with an overall mean of 2% missing item responses (range, 0.5-22.3%; Table 2). We found general support for the legitimacy of summing items without weighting or standardization, as illustrated by roughly similar item mean scores and SDs within most scales and corrected item-total correlations above the recommended criteria of 0.4 for all items (Table 2). All reliability coefficients exceeded the recommended minimum of 0.70, and all but five exceeded the preferred value of 0.80. However, the minimum reliability criterion of 0.7 was not reached in four instances (three scales) when taking the 95% confidence intervals into account (Table 3).

Table 2. Descriptive PDQ-39 scale and item statistics [a]

Scale / Item		Missing		Score [b]		Item-total correlation [c]
No.	Item problem area (abridged)	n	%	Mean (SD)	Median (q1, q3)	
Mobility (MOB)		**10**	**5**	**42.95 (28.43)**	**45 (20, 62.5)**	-
1	Leisure activites	2	1	2.03 (1.23)	2 (1, 3)	0.731
2	Looking after home	6	3	1.85 (1.33)	2 (1, 3)	0.818
3	Carry shopping bags	4	2	1.93 (1.50)	2 (0, 3)	0.787
4	Walking half a mile	3	1.5	1.95 (1.50)	2 (0, 3)	0.809
5	Walking 100 yards	6	3	1.25 (1.35)	1 (0, 2)	0.775
6	Getting around the house	5	2.5	1.73 (1.29)	2 (0, 3)	0.818
7	Getting around in public	4	2	1.92 (1.35)	2 (1, 3)	0.894
8	Need company when going out	4	2	1.56 (1.49)	1 (0, 3)	0.774
9	Worry falling in public	4	2	1.40 (1.30)	1 (0, 2)	0.704
10	Confined to the house	1	0.5	1.68 (1.24)	2 (0, 3)	0.808
Activities of daily living (ADL)		**3**	**1.5**	**38.94 (24.76)**	**37.5 (20.8, 58)**	-
11	Washing	1	0.5	1.07 (1.21)	1 (0, 2)	0.753
12	Dressing	2	1	1.43 (1.27)	1.5 (0, 2)	0.792
13	Do buttons or shoe laces	2	1	1.91 (1.26)	2 (1, 3)	0.767
14	Writing clearly	1	0.5	2.15 (1.20)	2 (1, 3)	0.636
15	Cutting food	1	0.5	1.62 (1.24)	2 (1, 3)	0.743
16	Hold a drink without spilling	2	1	1.20 (1.17)	1 (0, 2)	0.586
Emotional well-being (EMO)		**5**	**2.5**	**37.92 (21.05)**	**37.5 (20.8, 54)**	-
17	Depressed	2	1	1.85 (1.07)	2 (1, 3)	0.798
18	Isolated & lonely	4	2	1.26 (1.10)	1 (0, 2)	0.680
19	Weepy or tearful	3	1.5	1.25 (1.01)	1 (0, 2)	0.671
20	Angry or bitter	3	1.5	1.26 (0.99)	1 (0, 2)	0.678
21	Anxious	2	1	1.71 (1.0)	2 (1, 2)	0.751
22	Worried about the future	3	1.5	1.82 (1.07)	2 (1, 3)	0.709
Stigma (STI)		**5**	**2.5**	**27.54 (23.17)**	**25 (6.2, 43.8)**	-
23	Felt need to conceal PD	2	1	0.99 (1.13)	1 (0, 2)	0.660
24	Avoid eating/drinking in public	4	2	1.28 (1.17)	1 (0, 2)	0.616
25	Embarrassed due to PD	2	1	1.16 (1.15)	1 (0, 2)	0.779
26	Worried people's reactions	2	1	1.02 (1.0)	1 (0, 2)	0.693
Social support (SOC)		**47**	**23.3**	**14.78 (18.08)**	**8.3 (0, 25)**	-
27	Close relationships	4	2	0.67 (0.86)	0 (0, 1)	0.413
28	Support from partner	45	22.3	0.56 (0.93)	0 (0, 1)	0.654
29	Support from family or friends	6	3	0.64 (0.90)	0 (0, 1)	0.661
Cognitions (COG)		**6**	**3**	**33.03 (20.35)**	**31.2 (18.8, 50)**	-
30	Unexpectedly fallen asleep	2	1	1.19 (1.12)	1 (0, 2)	0.464
31	Concentration	5	2.5	1.46 (1.11)	2 (0, 2)	0.645
32	Poor memory	2	1	1.55 (1.06)	2 (1, 2)	0.525
33	Dreams or hallucinations	2	1	1.12 (1.08)	1 (0, 2)	0.480
Communication (COM)		**4**	**2**	**27.99 (24.19)**	**25 (6.2, 41.7)**	-
34	Speech	2	1	1.41 (1.20)	1 (0, 2)	0.799
35	Unable communicate properly	2	1	1.33 (1.16)	1 (0, 2)	0.870
36	Felt ignored	2	1	0.65 (0.87)	0 (0, 1)	0.627
Bodily discomfort (BOD)		**4**	**2**	**40.91 (24.07)**	**41.7 (25, 58.3)**	-
37	Painful cramps or spasms	2	1	1.38 (1.24)	1 (0, 2.75)	0.591
38	Pain in joints or body	3	1.5	1.90 (1.19)	2 (1, 3)	0.583
39	Unpleasantly hot or cold	3	1.5	1.63 (1.16)	2 (1, 2)	0.465

[a] Scale level data are bold.
[b] Scale scores can range between 0-100 (100 = maximum level of problems); item scores can range between 0-4 (0 = never; 1 = seldom; 2 = sometimes; 3 = often; 4 = always, or cannot do at all).
[c] Corrected for overlap.
SD, standard deviation; q1, first quartile (25th percentile); q3, third quartile (75th percentile).

Dimensionality

We found indications challenging whether the eight PDQ-39 scales represent the best grouping of items. Scaling success rates averaged 56.2% and did not reach 100% for any of the scales (Table 3). Only one of the eight PDQ-39 scales (SOC) showed signs (9.5%) of scaling failure.

Table 3. Reliability, scaling success and floor/ceiling effects of the PDQ-39

	Reliability			
	Cronbach's alpha [a] (95% CI)	Test-retest [b] (95% CI)	Scaling success (%) [a,c]	Floor / ceiling effect (%) [a,d]
MOB	0.95 (0.94-0.96)	0.93 (0.91-0.95)	75.7	11.4 / 2.5
ADL	0.89 (0.87-0.91)	0.93 (0.90-0.95)	59.5	7.9 / 1.0
EMO	0.89 (0.87-0.91)	0.87 (0.82-0.91)	57.1	5.4 / 0.5
STI	0.85 (0.81-0.88)	0.85 (0.79-0.89)	78.6	20.3 / 0.5
SOC	0.74 (0.66-0.81)	0.76 (0.66-0.83)	57.1	35.6 / 0 [e]
COG	0.74 (0.67-0.79)	0.86 (0.81-0.90)	21.4	6.9 / 0 [f]
COM	0.87 (0.83-0.90)	0.86 (0.81-0.90)	61.9	24.3 / 0.5
BOD	0.72 (0.65-0.78)	0.79 (0.72-0.85)	38.1	7.9 / 0.5

[a] From first administration.
[b] One-way random intra-class correlation calculated from scores of patients completing both administrations (2 weeks apart) themselves and reporting unchanged health at second administration (n=137).
[c] Percentage of occasions when items correlated significantly stronger with their proposed scale than with other scales.
[d] Percentage of sample scoring 0 (floor) and 100 (ceiling).
[e] Maximum observed score for SOC was 67.67.
[f] Maximum observed score for COG was 81.25.
CI, confidence interval; MOB, mobility; ADL, activities of daily living; EMO, emotional well-being; STI, stigma; SOC, social support; COG, cognitions; COM, communication; BOD, bodily discomfort.

Exploratory factor analysis yielded eight factors according to the criterion used by Peto et al.[7] However, the grouping of items did not accord with the assumed PDQ-39 scales and eigenvalues of several factors only marginally exceeded 1 (Fig. 1). Parallel analysis identified four factors that were stronger than those produced by random data (Fig. 1). Among these first four factors, two of the proposed scales (EMO and COM) were intact (factors 2 and 4, respectively). Factor 1 consisted of the ten MOB items and four ADL items, and factor 3 included the four STI items and one SOC item (Fig. 1). Confirmatory factor analysis showed poor fit (χ^2, 1885.85; P<0.0001) of the observed data to the proposed items-to-scales relationships, thus arguing against the assumed structure (see Supplementary Fig. S1 online for details).

- Figure 1 about here -

Rasch analyses revealed four scales (MOB, ADL, SOC and COM) with signs of overall lack of fit (χ^2, 16.7-41.0; P≤0.01) to the measurement model (see Supplementary Table S1 online for details). Individual item fit to the respective scales are reported in Table 4. A total of nine items, representing all scales but EMO, displayed signs of misfit. This suggests that these items do not work in harmony with the other items in their respective scales. Assessment of DIF identified significant DIF by gender for items 1 (MOB), 19 (EMO) and 24 (STI), and by age for item 24 (STI) (for examples, see Supplementary Fig. S2 online).

Table 4. Rasch item and fit statistics for the PDQ-39 [a]

		Item statistics [b]		Fit statistics		
	Item	Location	SE	Residual [c]	Chi square [d,e]	F-statistic [e,f]
MOB	1	-0.51	0.10	2.02	4.17	1.67
	2	-0.26	0.10	-0.27	2.81	1.55
	3	-0.36	0.08	0.76	5.20	3.44
	4	-0.43	0.08	-0.08	1.36	1.05
	5	0.68	0.09	-0.55	7.45	4.77
	6	0.09	0.10	-0.17	1.58	1.00
	7	-0.24	0.10	**-3.06**	**12.36**	**13.02**
	8	0.15	0.08	-0.34	1.31	0.04
	9	0.53	0.09	2.26	3.50	1.64
	10	0.34	0.10	-0.30	1.30	0.74
ADL	11	0.79	0.09	-1.53	7.18	**6.06**
	12	0.22	0.09	-1.90	8.15	**7.20**
	13	-0.56	0.09	-0.66	4.16	2.98
	14	-0.89	0.09	1.59	6.82	3.54
	15	-0.07	0.09	-0.09	0.57	0.62
	16	0.51	0.09	**2.86**	**12.41**	4.87
EMO	17	-0.91	0.11	-1.73	3.99	3.30
	18	0.18	0.10	0.62	0.68	0.27
	19	1.22	0.11	1.98	1.97	0.88
	20	0.49	0.11	1.25	0.17	0.08
	21	-0.43	0.11	-0.57	2.83	1.81
	22	-0.55	0.11	0.68	0.44	0.25
STI	23	0.05	0.10	0.56	0.92	0.22
	24	-0.31	0.10	1.70	2.66	1.20
	25	-0.12	0.10	-1.10	7.77	**6.90**
	26	0.37	0.11	0.56	1.67	0.93
SOC	27	0.47	0.12	1.87	5.87	3.53
	28	-0.40	0.12	-0.87	7.06	**7.13**
	29	-0.07	0.11	-0.27	4.27	3.27
COG	30	0.59	0.08	1.65	0.75	0.27
	31	-0.59	0.09	-0.96	**11.68**	**9.99**
	32	-0.61	0.09	1.01	0.30	0.07
	33	0.60	0.09	1.11	1.32	0.64
COM	34	-1.03	0.12	-0.50	1.06	0.86
	35	-0.80	0.13	-2.25	8.17	**10.78**
	36	1.82	0.14	2.31	7.50	3.29
BOD	37	0.41	0.08	-0.19	7.15	**5.59**
	38	-0.42	0.08	-0.22	3.60	3.15
	39	0.00	0.08	1.34	0.04	0.02

[a] Performed with the sample divided into three class intervals according to person locations on the measured variables. For details, see Refs 13, 14 and 30.
[b] Expressed in linear log-odds units (logits), with mean item location set at 0 for each scale.
[c] Log residuals summarize the deviation of observed from expected responses. Deviation from the recommended range of -2.5 to +2.5,[30] indicating item misfit, are bold.
[d] Chi square values summarize the deviation of observed from expected responses across the three class intervals of the sample. Higher absolute chi square values represent larger deviations.
[e] Bonferroni corrected statistically significant deviations across class intervals, indicating item misfit, are bold.
[f] One-way ANOVAs of deviations from model expectation across the three class intervals of people.
SE, standard error; MOB, mobility; ADL, activities of daily living; EMO, emotional well-being; STI, stigma; SOC, social support; COG, cognitions; COM, communication; BOD, bodily discomfort.

Targeting

Ceiling effects were absent or negligible whereas all scales displayed floor effects (mean across the eight scales, 15%) and three scales exceeded the recommended maximum of 15% (Table 3). This pattern became particularly evident in the Rasch analyses of the relationship between the distributions of persons relative to items. All scales thus tended to measure at a level corresponding to more severe health problems than that experienced by the sample (Fig. 2A). Figure 2B exemplifies this pattern for the EMO scale by displaying the distributions of person and item locations on their common logit metric. Superimposed on the person distribution graph is the information function curve (Fig. 2B). This curve can be interpreted as an inverse of the standard error of measurement and indicates at what locations people are measured with good precision and little error. In addition, as illustrated in Figure 2B and by the item locations in Table 4, items within each scale tended to represent a relatively narrow range of health problems.

- Figure 2 about here -

DISCUSSION

This study assessed the measurement assumptions and properties of the PDQ-39 using traditional and Rasch measurement methods. Because study design cannot compensate for ambiguous measurement properties,[25] such assessments are essential to guide use and interpretation of scales in clinical research. We found generally good data quality and reliability, as well as general support for the legitimacy of summing PDQ-39 items without weighting or standardization within the respective scales. However, violations of the assumption of unidimensionality, which is a fundamental requirement for summed rating scales, argue against the validity of summing PDQ-39 items into their suggested scales. All PDQ-39 scales exhibited a relative measurement bias towards more severe health problems. These results have implications for the role of the PDQ-39 in evidence-based medicine, as well as for future developments towards improved outcome measurement in PD. This is discussed below together with some possible explanations for the current observations.

Score reliability of the eight PDQ-39 scales was found acceptable, although it was suboptimal for three scales (SOC, COG and BOD). While this is encouraging, investigators should be aware that reliability is central in planning clinical studies, particularly when using rating scales as clinical trial endpoints. Compromised reliability, even if exceeding the minimal acceptable criteria, adversely impacts sample size requirements and needs to be taken into account since power calculations do not assume any measurement error.[25]

Whereas reliability is fundamental to evidence-based measurement, it does not tell what scores represent. This is a matter of validity, to which scale dimensionality is central. We found that it is unclear what the eight PDQ-39 scales represent and that they therefore should be interpreted with caution. While this appears to be the first independent study to assess the assumed grouping of PDQ-39 items with a sample size that is reasonable for, e.g., factor analysis,[28] our results largely agree with previous observations. For example, Tsang et al.[12] found an average scaling success rate of 58.6%; authors using exploratory factor analyses have failed to reproduce the eight assumed PDQ-39 scales;[9 11] and our own initial observations suggested deviations from unidimensionality in four PDQ-39 scales.[10] Ambiguous meaning of scores is considered a main limitation of currently available health status questionnaires in PD,[4] and clear support regarding what scores represent is now called for in order to support claims based on patient reported outcomes in clinical trials.[5] Available evidence suggests that it is unlikely that the eight PDQ-39 scales can be considered to meet such requirements. The apparent instability of the assumed PDQ-39 dimensionality may

relate to the reliance on exploratory factor analysis to select and group items into scales when the instrument was developed.[7] In addition to the tendency of the eigenvalue >1 criterion to overestimate the number of factors (scales),[26] item level exploratory factor analysis tends to produce spurious factors that reflect endorsement patterns rather than dimensionality. That is, items tend to cluster together due to their distributional properties even if they measure the same construct as other items.[24] Future scale developments would probably benefit from applying the Rasch measurement framework instead as this approach is not based on correlations and requires conceptualization of the measured constructs.[14-16]

Analyses of targeting suggest that the PDQ-39 does not conceptualise health problems at a level that is congruent with that experienced by people with PD. This became particularly evident in the Rasch analyses of the person-item distributions. As targeting relates to the characteristics of the investigated sample, our observations could be due to sampling effects. However, the people studied here represented a wide range of disease severity and duration, and their characteristics and PDQ-39 scores were similar to those previously reported from community-based and randomized samples.[33 34] Our observations regarding floor effects are also in general agreement with previous reports.[12 35-37] The levels of health problems that items represent relate to their contents. In addition to the use of exploratory factor analysis to select items (see above), targeting problems may therefore reflect characteristics of the people surveyed to generate and select the PDQ-39 items. However, no clinical information (e.g., stages or duration of PD) has been reported for the sample originally interviewed to generate PDQ-39 items.[7]

In addition to a general bias towards more severe problems we also found relatively narrow Rasch derived item locations, indicating that items represent fairly comparable levels of health problems. Similar observations were made by Ito et al.,[38] who failed in their attempt to develop PDQ-39 short forms targeted to different levels of PD severity because items covered very similar ranges. As a consequence of suboptimal targeting and clustering of items in the PDQ-39, and the relatively small number of items in several scales,[14 39] a considerable proportion of people are measured with relatively low degrees of confidence. This poses some limitations on the PDQ-39, particularly for clinical trials aimed to detect small but clinically important effects among people with less severe problems. For example, a recent randomized double-blind clinical trial comparing levodopa + entacapone with levodopa alone in mild to moderate PD found inconsistent results.[40] While clinician-reported motor and ADL scores favoured the levodopa + entacapone group, no differences were detected by PDQ-39 scales assumed to tap the same or similar constructs. This may, at least in part, have been due to suboptimal targeting and measurement precision of the PDQ-39.[40]

The findings reported here could be due to cultural differences or deficiencies with the Swedish version of the PDQ-39. However, there are reasons to believe that these are not major explanations. First, many of the issues identified here have been implied also in previous studies from various countries (see above). Second, the Swedish PDQ-39 has been carefully evaluated regarding linguistic validity.[18 19] However, empirical studies are needed to address these possibilities. Particularly, studies addressing the presence of DIF by languages/countries are warranted to assess the validity of pooling and comparing PDQ-39 data in international clinical trials.[13] Our sample may also pose some limitations to the generalizability of results. However, the primary purpose of the study was not to provide PDQ-39 scores representative for the general PD population, but to assess its measurement properties. Importantly, the sample represented a wide range of disease severity, duration and ages, and the distribution of most PDQ-39 scale scores spanned the full 0-100 range. There are also reasons to believe that our sample was fairly representative, given similarities with

previously reported international population based studies using the PDQ-39 (see above).[33 34] However, some subgroups (e.g., the oldest and most severely disabled) are probably under represented. Furthermore, this study has not assessed the PDQ-39 summary index or its 8-item short form, PDQ-8. These will need to be thoroughly assessed in separate studies, preferably by methods such as those used here as this appears to be lacking. Finally, a number of PD-specific health status questionnaires are currently available. While the PDQ-39 appears to be the most widely accessible and well documented alternative,[8] this study does not provide any information on its relative merits compared to other available instruments. As such studies currently appear to be lacking, comprehensive head-to-head psychometric comparisons are warranted to help determine the best available alternative for a given situation.

Our observations bear a number of implications to guide the use of the PDQ-39. While the eight scale scores appear reliable, clinicians should be aware that score interpretations are hampered by ambiguities regarding their meaning. Our observations suggest that the assumed eight-dimensional PDQ-39 structure may be overly complex (i.e., too many scales with too few items per scale). This is not only likely to impact the meaning of scores, but may also compromise other measurement properties adversely.[3 14 39] One remedy could be to redefine the questionnaire according to a more readily understood theoretical framework, for example by linking items to domains of the International Classification of Functioning, Disability and Health (ICF).[41] Techniques for doing this have recently been proposed and results from linking generic scales to the ICF have shown promise.[42] Such work may not only help improving interpretation of scores but also, in combination with quantitative techniques such as Rasch analysis, provide a basis for item reduction, which could lessen respondent burden.[19]

Caution should be exercised when interpreting PDQ-39 trial data that fail to detect differences or changes over time (particularly improvements), since compromised responsiveness is a likely consequence of suboptimal targeting and measurement precision. In order to rectify this, new items that conceptualize less severe problems are probably needed. Indeed, expanding the item pool could serve both to increase measurement precision and to decrease respondent burden, if conducted by means of so called item banking.[3 14 43] This technique allows for selection of study specific, or even personally tailored, subsets of items without substantial loss of measurement precision or validity.[44]

The PDQ-39 has made, and will continue to make, significant contributions to our understanding of the impact of PD. However, this does not preclude seeking to improve the scale. Rating scale properties are relative and their adequacy relate, in part, to the purpose and context of their use. In this study the eight PDQ-39 scales were assessed primarily in perspective of their use as clinical trial endpoints. Unambiguous and valid inferences regarding the effectiveness of treatments require high quality outcome measures that meet rigorous scientific standards.[3-6 14] Our observations suggest that the ability of the PDQ-39 to meet such standards can be challenged. In order to further clarify the role of the PDQ-39, we encourage others to examine their data and recommend that measurement properties should be reported in studies using PDQ-39 endpoints.

Acknowledgements The authors wish to thank all participating patients for their cooperation, Jan Reimer for assistance with data collection and Elisabeth Rasmusson for secretarial assistance.

Competing interests None.

Funding The study was supported by the Swedish Research Council, the Skane County Council Research and Development Foundation, Rådet för hälso- och sjukvårdsforskning (HSF), and the Department of Nursing. C.N. was supported by the Section of Occupational Therapy & Gerontology, Lund University.

Legends to Figures

Fig. 1.

Scree plot of the eigenvalues (y-axis) for factors (x-axis) identified by item level exploratory principal component factor analysis of the PDQ-39 and 1000 parallel sets of randomly generated PDQ-39 data. Plots represent PDQ-39 eigenvalues (grey squares, empirical data) and the 95th percentile of 1000 random data eigenvalues (black diamonds). The dashed horizontal line indicates the cut point for determination of the number of factors (scales) according to the eigenvalue >1 criterion.[7] This criterion identified eight factors (Kaiser-Meyer-Olkin measure of sampling adequacy: 0.92; Bartlett's test of sphericity: χ^2, 4390.2, P<0.0001), of which the first four were stronger than those produced by random data. Contents of these four factors are indicated in the Figure. The first four empirical and random factors explained 59.1% (PDQ-39) and 19.3% (random data) of the total variance. Factors five through eight explained an additional 12.5% (PDQ-39) and 16.1% (random data) of the total variance. MOB, mobility; ADL, activities of daily living; EMO, emotional well-being; STI, stigma; SOC, social support; COM, communication.

Fig. 2.

PDQ-39 scales' targeting of the sample as assessed by Rasch analyses. (A) Mean person locations relative to the mean item locations (set at 0 logits). The mean person location across the eight scales was -1.32 logits below the items. (B) Detailed example of targeting (for the EMO scale). Distributions of the locations of people and items on the common logit metric (negative values = better emotional well-being) are depicted on the upper and lower panels, respectively. Superimposed on the person distribution graph is the information function curve (higher values = less error and more information in scores, i.e., better measurement precision). The information function curve indicates that about half of the sample (to the left of the dashed vertical line) is measured with a relatively low degree of confidence. Reasonable information functions for the other scales were within ranges similar to that for the EMO scale, i.e., spanning approximately between -1.5/-1 to +1/+1.5 logits (data available on request). MOB, mobility; ADL, activities of daily living; EMO, emotional well-being; STI, stigma; SOC, social support; COG, cognitions; COM, communication; BOD, bodily discomfort.

REFERENCES

1. **Rascol O**, Goetz C, Koller W, Poewe W, Sampaio C. Treatment interventions for Parkinson's disease: an evidence based assessment. Lancet 2002; **359**: 1589-98.
2. **Wheatley K**, Stowe RL, Clarke CE, Hills RK, Williams AC, Gray R. Evaluating drug treatments for Parkinson's disease: how good are the trials? BMJ 2002; **324**: 1508-11.
3. **Hobart J**. Rating scales for neurologists. J Neurol Neurosurg Psychiatry 2003; **74**(Suppl IV): iv22-iv26.
4. **Marras C**, Lang AE. Outcome measures for clinical trials in Parkinson's disease: achievements and shortcomings. Expert Rev Neurother 2004; **4**: 985-93.
5. **Food and Drug Administration**. Draft Guidance for Industry. Patient-Reported Outcome measures: Use in Medicinal Product Development to Support Labeling Claims. Federal Register 71(23): 5862-5863. February 3, 2006. Available from: http://www.fda.gov/cder/guidance/5460dft.pdf
6. **Scientific Advisory Committee of the Medical Outcomes Trust**. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002; **11**: 193-205.
7. **Peto V**, Jenkinson C, Fitzpatrick R, Greenhall R. The development and validation of a short measure of functioning and well being for individuals with Parkinson's disease. Qual Life Res 1995; **4**: 241-8.
8. **Marinus J**, Ramaker C, van Hilten JJ, Stiggelbout AM. Health related quality of life in Parkinson's disease: a systematic review of disease specific instruments. J Neurol Neurosurg Psychiatry 2002; **72**: 241-8.
9. **Bushnell DM**, Martin ML. Quality of life and Parkinson's disease: translation and validation of the US Parkinson's Disease Questionnaire (PDQ-39). Qual Life Res 1999; **8**: 345-50.
10. **Hagell P**, Whalley D, McKenna SP, Lindvall O. Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham Health Profile. Mov Disord 2003; **18**: 773-83.
11. **Auquier P**, Sapin C, Ziegler M, Tison F, Destee A, Dubois B, et al. Validation en langue française d'un questionnaire de qualité de vie dans la maladie de Parkinson: le Parkinson's Disease Questionnaire - PDQ-39. Rev Neurol (Paris) 2002; **158**: 41-50.
12. **Tsang KL**, Chi I, Ho SL, Lou VW, Lee TM, Chu LW. Translation and validation of the standard Chinese version of PDQ-39: a quality-of-life measure for patients with Parkinson's disease. Mov Disord 2002; **17**: 1036-40.
13. **Tennant A**, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 2004; **42**: I37-48.
14. **Hobart JC**, Riazi A, Thompson AJ, Styles IM, Ingram W, Vickery PJ, et al. Getting the measure of spasticity in multiple sclerosis: the Multiple Sclerosis Spasticity Scale (MSSS-88). Brain 2006; **129**: 224-34.
15. **Tennant A**, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004;**7**(Suppl 1):S22-6.
16. **Wilson M**. Constructing measures: an item response modelling approach. Mahwah: Lawrence Erlbaum Associates, Inc., 2005.
17. **Gibb WRG**, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson’s disease. J Neurol Neurosurg Psychiatry 1988;**51**:745-52.
18. **Hagell P**, McKenna SP. International use of health status questionnaires in Parkinson's disease: translation is not enough. Parkinsonism Relat Disord 2003; **10**: 89-92.

19. **Kim MY**, Dahlberg A, Hagell P. Respondent burden and patient-perceived validity of the PDQ-39. Acta Neurol Scand 2006; **113**: 132-7.
20. **Hoehn MM**, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology 1967; **17**: 427-42.
21. **Likert RA**. A technique for the development of attitudes. Arch Psychol 1932; **140**: 5-55.
22. **Ware JE Jr.**, Harris WJ, Gandek B, Rogers BW, Reese PR. MAP-R for Windows: multitrait/multi-item analysis program - revised user's guide. Boston, MA: Health Assessment Lab., 1997.
23. **Cronbach LJ**. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297-334.
24. **Nunnally JC**, Bernstein IH. Psychometric theory. New York: McGraw-Hill, Inc., 1994.
25. **Fleiss JL**. Design and analysis of clinical experiments. New York: John Wiley & Sons, Ltd., 1986.
26. **Zwick WR**, Velicer WF. Comparison of five rules for determining the number of components to retain. Psychol Bull 1986; **99**: 432-42.
27. **O'Connor BP**. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. Behav Res Methods Instrum Comput 2000; **32**: 396-402.
28. **Floyd FJ**, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess 1995; **7**: 286-99.
29. **Rasch G**. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research; 1960 (Reprinted: Chicago, University of Chicago Press; 1980).
30. **Andrich D**, Sheridan B, Luo G. Interpreting RUMM 2020. Perth, WA: RUMM Laboratory Pty Ltd; 2004–2005. Available from: http://www.rummlab.com.au/
31. **McHorney CA**, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995; **4**: 293-307.
32. **Barnette JJ**. ScoreRel CI: an Excel program for computing confidence intervals for commonly used score reliability coefficients. Educ Psychol Meas 2005; **65**: 980-3.
33. **Schrag A**, Jahanshahi M, Quinn N. How does Parkinson's disease affect quality of life? A comparison with quality of life in the general population. Mov Disord 2000; **15**: 1112-8.
34. **Findley L**, the Global Parkinson's Disease Survey (GPDS) Steering Committee. Factors impacting on quality of life in Parkinson's disease: results from an international survey. Mov Disord 2001; **17**: 60-7.
35. **Tan LC**, Luo N, Nazri M, Li SC, Thumboo J. Validity and reliability of the PDQ-39 and the PDQ-8 in English-speaking Parkinson's disease patients in Singapore. Parkinsonism Relat Disord 2004; **10**: 493-9.
36. **Luo N**, Tan LC, Li SC, Soh LK, Thumboo J. Validity and reliability of the Chinese (Singapore) version of the Parkinson's Disease Questionnaire (PDQ-39). Qual Life Res 2005; **14**: 273-9.
37. **Jenkinson C**, Fitzpatrick R, Norquist J, Findley L, Hughes K. Cross-cultural evaluation of the Parkinson's Disease Questionnaire: tests of data quality, score reliability, response rate, and scaling assumptions in the United States, Canada, Japan, Italy, and Spain. J Clin Epidemiol 2003; **56**: 843-7.
38. **Ito Y**, Yamaguchi T, Ohashi Y, Obu S, Kondo T, Kohmoto J, Nagaoka M, Fukuhara S. Using item-response theory to select items from the PDQ-39 that match the severity of Parkinson's disease. Qual Life Res 2000; **9**: 1058.
39. **Wright BD**, Masters GN. Rating scale analysis. Chicago: MESA Press, 1982.
40. **Reichmann H**, Boas J, MacMahon D, Myllyla V, Hakala A, Reinikainen K. Efficacy of combining levodopa with entacapone on quality of life and activities of daily living in patients experiencing wearing-off type fluctuations. Acta Neurol Scand 2005; **111**: 21-8.

41. **WHO**. International classification of functioning, disability and health. Geneva: World Health Organization, 2001.
42. **Cieza A**, Stucki G. Content comparison of health-related quality of life (HRQOL) instruments based on the international classification of functioning, disability and health (ICF). Qual Life Res 2005; **14**: 1225-37.
43. **Bode RK**, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. Arch Phys Med Rehabil 2003;**84**(Suppl 2):S52-60.
44. **Ware JE, Jr.**, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000; **38**: II73-82.

Fig. 1

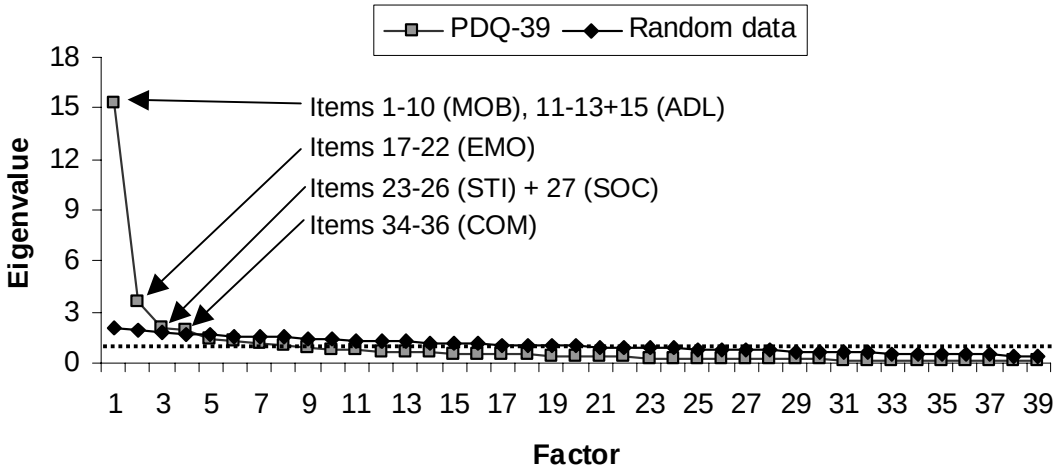
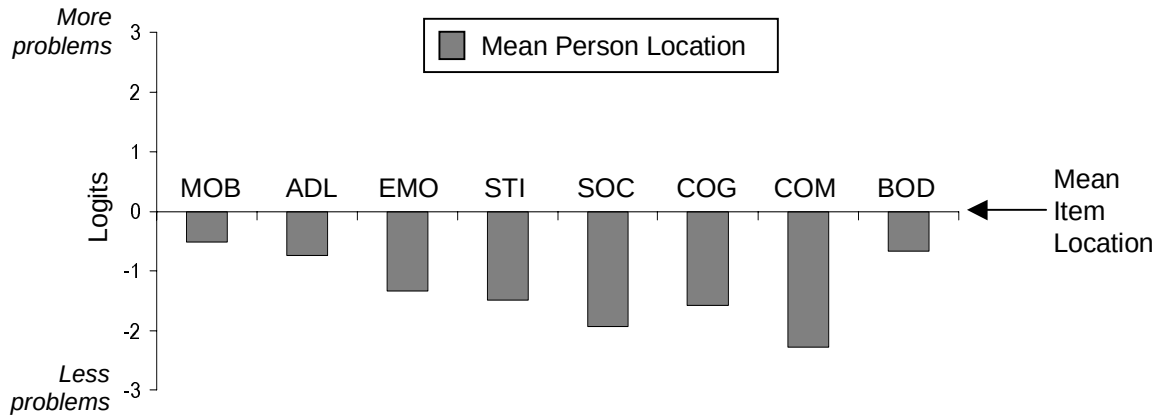
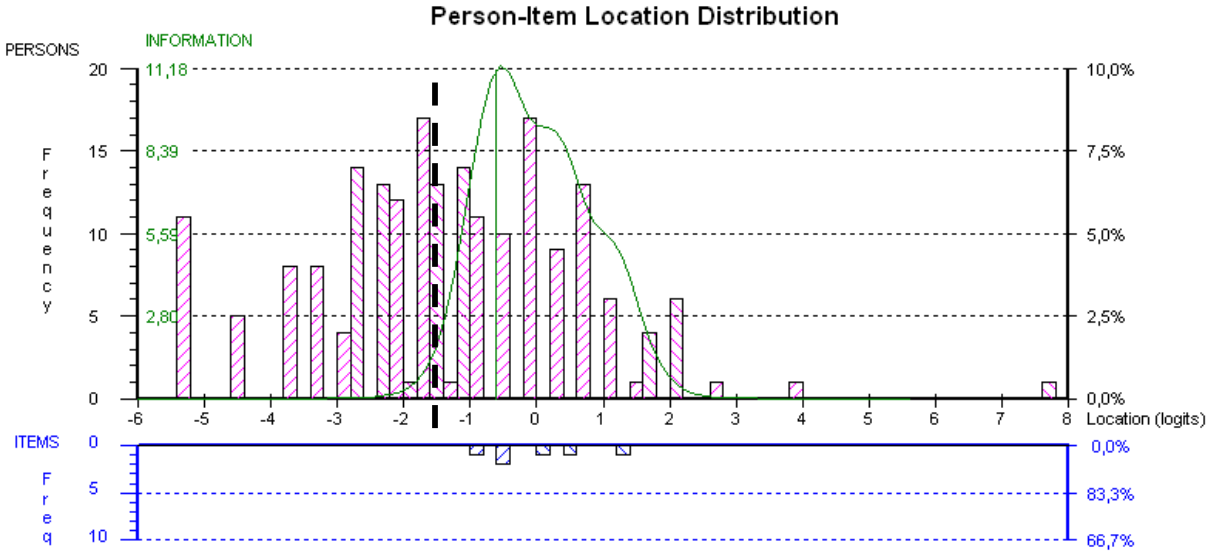


Fig. 2

A



B



Supplementary Table S1.
Overall Rasch model fit statistics and reliability of PDQ-39 scales [a]

	PDQ-39 scales							
Overall fit statistics	MOB	ADL	EMO	STI	SOC	COG	COM	BOD
Items								
Fit residual (mean) [b]	0.031	0.037	0.372	0.428	0.245	0.703	-0.148	0.310
Fit residual (SD) [c]	1.481	1.842	1.329	1.152	1.442	1.144	2.301	0.897
Persons								
Fit residual (mean) [b]	-0.350	-0.290	-0.379	-0.406	-0.555	-0.290	-0.591	-0.288
Fit residual (SD) [c]	1.266	1.006	1.098	1.230	1.281	1.088	1.336	0.849
Total item-trait interaction								
Total item chi-square	41.046	39.275	10.081	13.021	17.199	14.054	16.729	10.779
df	20	12	12	8	6	8	6	6
P-value	0.004	<0.001	0.609	0.111	0.008	0.080	0.010	0.095
Person separation index [d]	0.96	0.90	0.90	0.85	0.72	0.71	0.88	0.71

[a] As analysed using RUMM2020 (Rumm Laboratory Pty Ltd., Perth) for Windows.
[b] Should be close to 0 (Andrich *et al*., 2004-2005).
[c] Should be close to 1 (Andrich *et al*., 2004-2005).
[d] Rasch based reliability statistic (analogous to Cronbach's alpha).
SD, standard deviation; df, degrees of freedom; MOB, mobility; ADL, activities of daily living; EMO, emotional well-being; STI, stigma; SOC, social support; COG, cognitions; COM, communication; BOD, bodily discomfort.

Reference
Andrich D, Sheridan B, Luo G. Interpreting RUMM2020. Perth, WA: RUMM Laboratory Pty Ltd; 2004–2005.