Scales for rating motor impairment in Parkinson's disease: studies of reliability and convergent validity

L Henderson, C Kennard, T J Crawford, S Day, B S Everitt, S Goodrich, F Jones, D M Park

Abstract

Study 1 examined the reliability of the ratings assigned to the performance of five sign-and-symptom items drawn from tests of motor impairment in Parkinson's disease. Patients with Parkinson's disease of varying severity performed gait, rising from chair, and hand function items. Video recordings of these performances were rated by a large sample of experienced and inexperienced neurologists and by psychology undergraduates, using a four point scale. Inter-rater reliability was moderately high, being higher for gait than hand function items. Clinical experience proved to have no systematic effect on ratings or their reliability. The idiosyncracy of particular performances was a major source of unreliable ratings. Study 2 examined the intercorrelation of several standard rating scales, comprised of sign-and-symptom items as well as activities of daily living. The correlation between scales was high, ranging from 0·70 to 0·83, despite considerable differences in item composition. Inter-item correlations showed that the internal cohesion of the tests was high, especially for the self-care scale. Regression analysis showed that the relationship between the scales could be efficiently captured by a small selection of test items, allowing the construction of a much briefer test.

The advent of levodopa replacement therapy gave impetus to the development of clinical rating scales for assessing impairment in Parkinson's disease (see Marsden and Schachter,[1] Potvin and Tourtellotte[2] for reviews). Despite continuing proliferation of scales, few attempts have been made to evaluate their reliability or validity, or to provide a rationale for the selection of constituent items. Test items tend to fall into two broad categories, sign-and-symptom items which are essentially formalisations of the tests used in the consulting room to reveal Parkinsonian impairment, and Activities of Daily Living (ADL) items which assess the functional status of the patient in a more global fashion.

In Study 1, we report an investigation of the inter-rater reliability of sign-and-symptom items and of the factors that influence reliability. In Study 2, we examine the intercorrelation between different scales and between test items of different types, to determine how much redundancy the tests possess and to investigate the extent to which different types of test item converge on the same underlying properties.

STUDY 1

Aim

Our general intention in this study was to investigate some of the factors that should guide the selection of sign-and-symptom based test items. In particular, we were concerned with inter-rater reliability and how it might be improved. Factors that might be expected to influence the reliability of subjective ratings include the following: *Item selection*: the nature of the particular movement that is assessed, how familiar it is, how revealing of abnormality, how many biomechanical degrees of freedom it admits, etc. We investigated gait, rising from a chair and three hand function items. *Item standardisation*: we specified to the patient the action required, both verbally and by demonstration. *Rater expertise*: we compared ratings made by neurologists experienced in the use of Parkinsonian rating scales, inexperienced neurologists and psychology undergraduates. We also investigated the effects of a brief training video. *Rating criteria*: we provided either brief written criteria or video demonstrations of prototypic examples of each scale value. *Contextual factors*: these include simultaneous context (strictly irrelevant factors such as expressions of distress or tremor in limbs other than the one being assessed) and prior context (expectations derived from previous assessments of the patient or criterion bias induced by having just rated much more/less impaired patients). In this study some patients were rated several times, in different clinical states. An attempt was also made to vary the amount of simultaneous context available.

Methods

We began by making a video recording of 11 patients with idiopathic Parkinson's disease performing five test items. From this video data-base we selected and edited appropriate examples to produce our test video, which was used to gather ratings. A four point rating scale was employed for all test items. Ratings were obtained from 50 physicians, 44 of whom were neurologists, the remainder being geriatricians (henceforth, simply "neurologists"). We also obtained ratings from 80 psychology undergraduates.

The Department of Neurology, The London Hospital, Whitechapel, London
C Kennard
T J Crawford
S Goodrich
D M Park

Psychology Division, Hatfield Polytechnic, Hatfield,Hertfordshire
L Henderson
F Jones
S Day

Biometrics Unit, Institute of Psychiatry, Denmark Hill, London, United Kingdom
B S Everitt

Correspondence to:
Professor Henderson, Psychology Division, Hatfield Polytechnic, Hatfield, Hertfordshire AL10 9AB, United Kingdom.

Received 11 January 1990
and in revised form
3 May 1990.
Accepted 7 June 1990

The patients All the patients filmed were outpatients at the London Hospital who had consented to be videod and for the recordings to be used for scientific purposes. A few patients were filmed twice, once when medication had been delayed and again after medication, when their usual level of functioning had been restored.

The test recording This comprised two parts, the second of which was prefaced by a brief training demonstration. In both parts, five patients performed each of five items: rising from a chair, walking and turning, finger taps, finger flexion and wrist pronation/supination. (We reserve the term "performance" for the recording of an individual patient's performance of a test item.) The two parts of the film differed in the following respects. In part 1, performances of all five items were recorded in sequence, patient by patient. First came a demonstration run, with one patient performing each of the five test actions, in turn. These were not rated. Then came the series of five patients for rating, patients 2 and 5 being the same individual tested in two states of medication.

In part 2, there were again five different performances of each of the five test items, to be rated. However, the following changes were made from the format adopted in part 1. Rather than the various tests being presented patient by patient, the performances were presented item by item, with five different patients performing each test item before the next test item was encountered. Moreover, different (but overlapping) sets of patients were chosen for each test item. Before the five performances of each item a training demonstration was presented. This consisted of a prototypical example of a performance meriting each of the four ratings (0 = normal; 1 = mild; 2 = moderate; 3 = severe). These four examples were each viewed twice. The first of these presentations was accompanied by a commentary pointing out any abnormal features of the movement. Two other changes distinguished part 2 from part 1. The walking and turning item was not divided into separately rated armswing and gait ratings; instead, one single composite rating was given. Also, wherever possible, the face of the patient and irrelevant parts of the body were blanked out.

The test items The items were drawn with minor amendments from the Webster[3] and Unified Parkinson's Disease Rating Scale (UPDRS)[4] tests. Rising from a chair was performed from a hard, upright chair with arms. The patient was instructed to attempt initially to rise without using leverage on the arms of the chair. Gait involved walking at a natural pace for six metres, turning within the confines of a box marked out on the floor and returning to the starting point.

The three hand function items were performed seated. Finger tapping was performed with the index finger, while the hand was rested flat upon the table, palm down. Finger flexion and wrist pronation/supination were performed with the relevant arm extended straight in front, level with the shoulder. Finger flexion required the repetitive opposition of thumb and index finger. Pronation and supination were performed in alternation, with fingers partially extended. For these items, the 10 seconds of recorded activity were culled from the end of a 24 second performance.

Raters and rating procedure The fifty United Kingdom neurologists were tested together during a symposium on Parkinson's disease. For some of the subsequent analyses they were divided into experienced/inexperienced subgroups, according to whether they had declared a particular interest in motor disorders and had experience of using the Webster test (N = 23) or lacked at least one of these attributes, usually both (N = 27). The undergraduate raters were first year (PSY1) and second year (PSY2) psychology students. PSY2 (N = 38) were run as a single group. PSY1 were divided into two subgroups, who either rated the video in the standard order (Pt 1 − > Pt 2: N = 19) or in the reverse order (Pt 2 − > Pt 1: N = 23).

The rating criteria to be used for part 1 were provided for the raters on a printed sheet. The criteria used in part 2 were provided by the training examples.

Statistical analysis The four-point ratings (0–3) for each item were used to calculate two measures of the agreement among groups, the standard deviation and the coefficient of concordance. The standard deviation was our primary index and provided a measure of the variability across a group of the ratings assigned to a particular performance. Kendall's Concordances[5] have also been calculated to provide a measure of the agreement between raters as to the ranking of patients. We cite concordances to allow comparison with previous reliability studies.[4 6 7] However, as will appear later, a problem with this coefficient is that it may be biased by the degree of similarity found in a particular set of patients.

Results

Reliability and expertise Figure 1 shows the mean ratings given by the neurologists (undivided) and the undergraduates to each of the 30 rateable items (five patients × six items) in part 1. Both groups seem to use the full range of the scale and their mean ratings agree closely, showing similar profiles for each patient.

Figure 2 displays the inter-rater variability (SD) of the ratings assigned to each item performance in part 1, as a function of the mean level of impairment indicated by the ratings. Only the ratings by neurologists are shown but a very similar inverted-U function was found for undergraduate ratings. What this pattern of results indicates is that the only systematic relationship between a patient's mean rated level of impairment on a test item and the inter-rater variability of these ratings, occurs at the extreme ends of the scale (< 0·25 and > 2·75), where variability declines sharply.

Figure 3 allows inspection of the average rating variability of each test item, as a function of expertise. These data, trimmed of items with

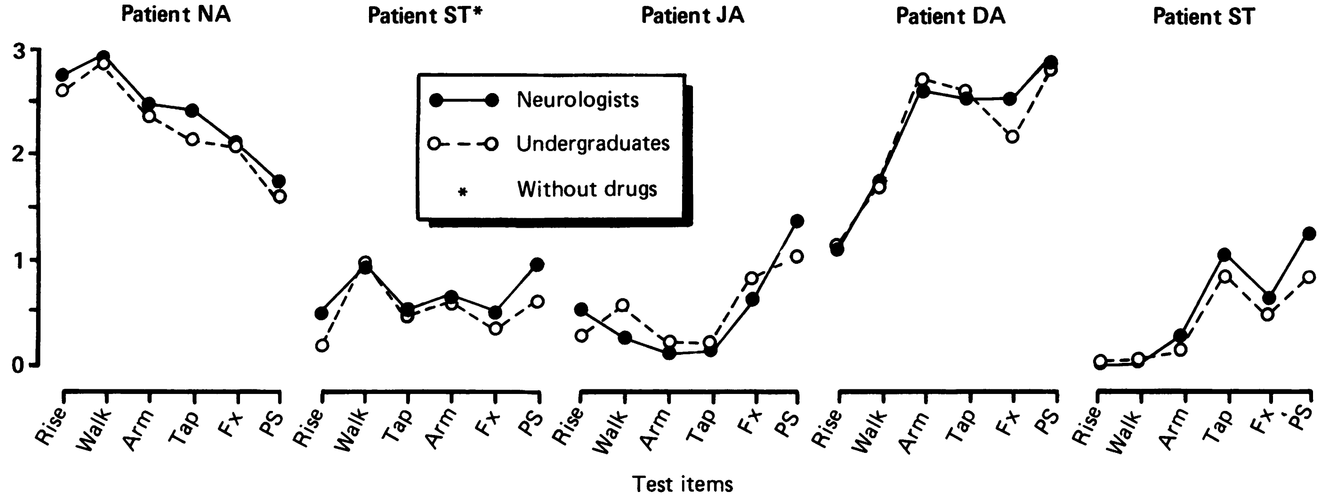
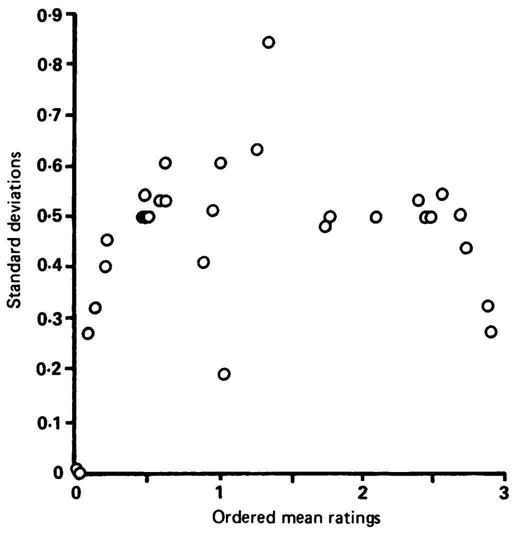


Figure 1 Mean ratings assigned to performances in Part 1 by neurologists and undergraduates. (Test items were rise from chair, walking and turning, finger tapping, finger flexion and wrist pronation/supination).

Figure 2 Variability (SD) of the rating assigned by the neurologists to a performance, as a function of its mean rated impairment (Part 1).



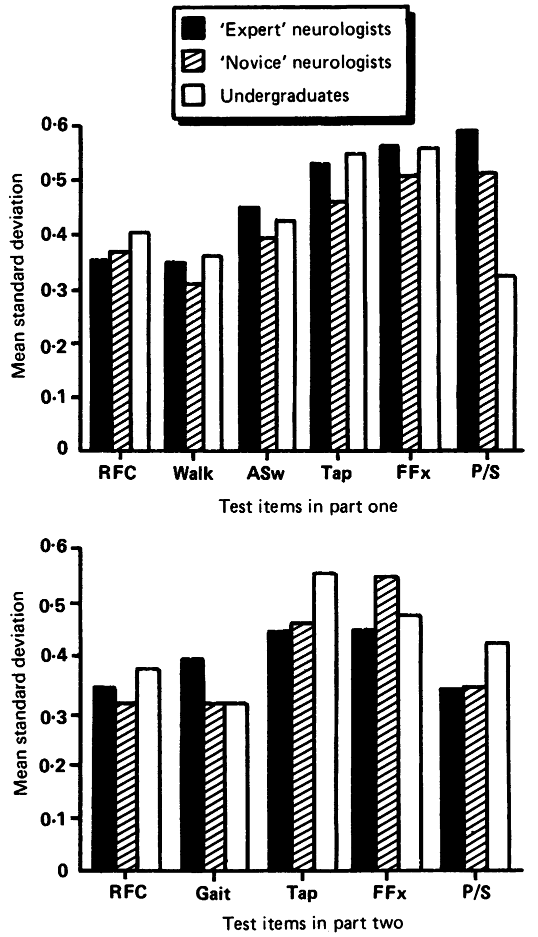


Figure 3 Variability (SD) as a function of rater experience and item type, shown separately for Parts 1 and 2.

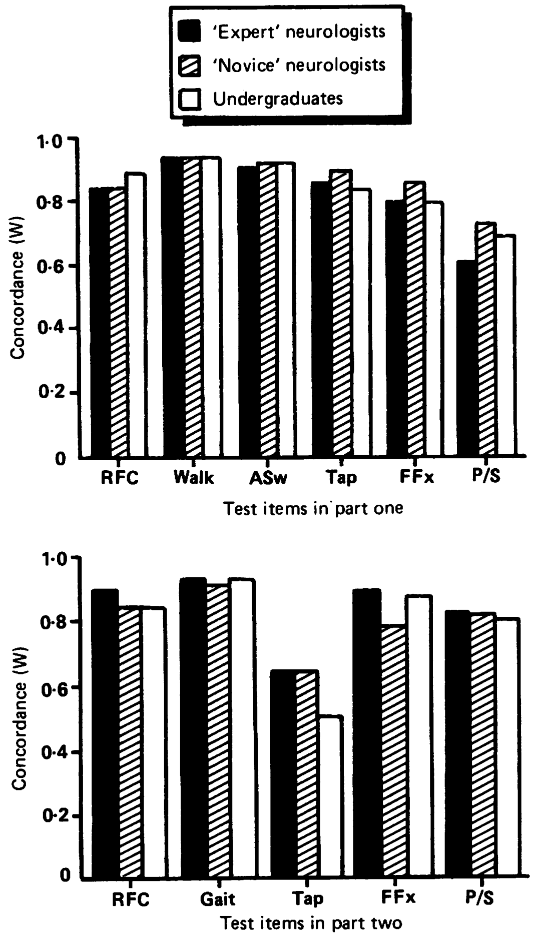
a mean above 2·75 or below 0·25, were submitted to a three-way ANOVA (groups × parts × items). While, overall, variability was greatest for the undergraduates and least for the inexperienced neurologists, this expertise factor did not approach significance. Variability was reduced in part 2 and this effect was just significant ($p < 0.05$). The effect of item type was highly significant ($p < 0.001$) with rising from chair and gait both showing higher reliability (mean SDs 0·34) than the hand function items (SDs 0·48–0·52).

Figure 4 displays the concordance coefficients for each test item in parts 1 and 2. (Note that high scores now denote agreement.) Here, also, there is no indication that experts' ratings yield more agreement but there is a tendency in all the groups for hand function items to yield less agreement. Finger taps in part 2 produced notably low concordance but subsequent analyses showed this to be due to a reduced range of impairment on this item.

Training To evaluate the effect of the training demonstration that introduced each test item in part 2, we compared the reliability measures obtained in part 1 for the two PSY1 subgroups. The reverse order subgroup, who alone had seen the training examples by this point, showed no resulting benefit, indicating that our brief training with exemplars had been ineffective as a means of improving agreement amongst inexperienced raters.

Qualitative observations Item standardisation presents a major problem, especially for

Figure 4 Concordance within groups for each test item, shown separately for Parts 1 and 2.



the hand function items used here. The actions required are not everyday movements and despite clear instructions/demonstrations patients found a variety of ways to execute the movements. Postural idiosyncracies may affect the ease of execution. Such variation also complicates the rater's task. (A solution that we are currently exploring is to employ a device that constrains the movement, such as a rotating handle for pronation/supination.) Another source of difficulty in rating alternating hand movements is that with severely impaired patients, unless instructions emphasise movement amplitude rather than rate, tremor may be taken for rapidly alternating movements of low amplitude. Indeed, we suspect that some patients generate tremor as a movement surrogate.

Two types of contextual cue may obstruct the attempt to focus on a specific feature and rate it independently. The first of these we believe to be responsible for the high SD outlier evident in fig 2. (JA performing P/S. Note that this outlier is not a "capricious" datum. This performance also yielded the largest SD for the undergraduates.) While the pronation/supination is itself only very mildly impaired, the patient's face shows intense effort and there is some postural tremor. Different weightings attached to these concurrently available cues produce unreliable ratings. As it happens, JA is also the subject of sequential context cues, since she appears several times throughout the video, in very different states of medication.

Finally, another reason for supposing that the idiosyncracies of particular performances represent a major source of variability in the ratings is that the pattern of SD obtained across performances was remarkably similar for the neurologists and undergraduates ($r = 0{\cdot}659$; $p < 0{\cdot}001$).

We conclude that careful selection of test items, standardisation of their manner of execution, the clarification of rating criteria and removal of contextual cues seems more likely to improve reliability than does the selection of raters on the basis of experience or the provision of very brief training examples.

STUDY 2

In this study, we drew upon data gathered in the course of a double-blind drug investigation.[9] This data-base comprised the scores of 49 patients on five different tests of impairment. The tests represent the full range of item types, some being ADL-based and others being sign-based. These data allowed us to pursue two main questions: 1) Does the relationship between patients' scores on different scales offer persuasive evidence of convergent validity? (Where an external, objective validating criterion or "gold-standard" is unavailable, weaker evidence of the validity of a measuring instrument can be found in its tendency to agree with other tests that purport to measure the same features. Convergent evidence of validity is more impressive the more the content of the test items differs); 2) Is there sufficient redundancy amongst the test items to allow construction of a much briefer test that nevertheless correlates well with existing instruments.

Patients The sample was screened to exclude patients with dementia or with an additional condition that might contribute to the assessed impairment.

Test The scales used were: 1) Northwestern University Disability Scale (NUDS),[6] comprising five ADL items (walking, dressing, eating/feeding, hygiene, speech), each rated on a 10 point scale, save eating/feeding which both have a five point scale. Total possible score = 50. (50 = normality); 2) Hoehn and Yahr staging,[10] categorises patients into five stages, using multiple criteria. (0 = normality); 3) Self Care Scale[9]—self-ratings of 12 items (dressing, eating, food preparation, house cleaning, getting out of bed, turning in bed, rising from chair, climbing stairs, use of toilet, use of tools, bathing, shopping/mobility), each rated on a four point scale (0–3). Total possible score = 36. (0 = normality); 4) Webster Scale,[2] largely a sign-based test containing 10 items (manual bradykinesia, rigidity, posture, arm swing while walking, gait, tremor, facies, seborrhoea, speech, self-care), each rated on a four point scale (0–3). Total possible score = 36. (0 = normality); 5) Karnofsky Performance Score,[11] originally devised to provide a classification of cancer patients' functional self sufficiency into 10 gradations (from 100 = "normal" to 0 = "dead").

Results and discussion

Intercorrelation of the composite scores Table 1 shows some of the intercorrelations of the composite test scores. All the correlations are quite high, accounting for at least 50% of the variance. This is not entirely surprising, given the overlap of content between the tests. Most interesting therefore is the high correlation between the Webster and NUDS (0·82), since these scales overlap negligibly in content. In fact, the correlation of the NUDS with the Webster (different item types) is as great as that between the NUDS and Self Care (0·81) both of which are ADL-based. Brown *et al*[12] also found a similarly high correlation between scores on an ADL scale and a symptom-based scale (the King's College Hospital Parkinson's disease rating scale). In their study, agreement was greater when the ADL ratings were made by the clinician who had rated the symptom items, rather than when ADL scores derived from patients' self ratings.

ADL versus sign-based items To examine the relation between ADL and sign-based items, the intercorrelations between all the constituent items in the Webster and NUDS were calculated (table 1). Of the sixty correlations between individual items, only five failed to attain significance ($p < 0·05$). Three of these involved speech items.

Hoehn and Yahr No patient was assigned to stage 5. Those patients classified as stage 4 revealed, as a group, reliably greater impairment scores than those at stage 3 on all the other measures except the Webster. However, classification into stages 1–3 showed a very inconsistent relationship with the NUDS, Self Care and Webster scores. We therefore pursued the relationship between Hoehn and Yahr, and Webster scores in a larger data-base, acquired by the United Kingdom Parkinson's Disease Research Group (PDRG) and comprising 175 untreated patients, at stages 1–3. The groups of patients rated as stage 1 and 2 could not be reliably distinguished on their Webster scores. Taken together with other data suggesting that Hoehn and Yahr scores may behave anomalously,[13] and that the reliability of the measure may be low,[8] this result suggests that the ease of arriving at Hoehn and Yahr stagings is won at excessive cost, at least for the earlier stages.

Redundancy and test length Larsen *et al*[14] have proposed that a test of Parkinsonism should be brief, to avoid fatigue and to permit multiple testing during a patient's day. Given the relatively high intercorrelations we obtained between test items, it seemed worth investigating whether a much briefer instrument could be constructed as an index of the functional competences on which the test items tend to converge. To this end, the NUDS total scores were regressed on the 22 items which comprised the Webster and Self Care scales. The seven test items showing the largest regression coefficients relative to their standard errors are shown in table 2. The new scale comprising these items correlates with the NUDS at 0·89 and with the Karnofsky at 0·82, suggesting considerable convergence, at a cost of relatively few items in the new scale.

Since the most remarkable recent development in measures for rating Parkinsonism has been the construction of the chimerical UPDRS,[4] a vast instrument assembled from components of several existing scales, it seems worth posing the question as to what might be gained by grossly extending the number of items in a scale. Given the coarse, four-point scale employed by both the Webster and the Self Care Inventory, it seems unlikely that increasing the number of items has the potential greatly to enhance sensitivity of the scale. However, a modest degree of item redundancy may be justifiable, in terms of a consequent improvement in reliability of the total score. To achieve finer discriminations, at least one of the following steps must be taken: 1) A coarse scale may be retained but item difficulty varied, with the result that items differ in the portion of the continuum of disability that they resolve. This is essentially the technique employed by most intellectual tests, where each item produces only a binary (right/wrong) score; 2) An attempt can be made to transcend the normal limitation of a raters' capacity for absolute judgements by anchoring scale values in detailed descriptions, such as the NUDS provides for its ten-point scales; 3) Change of state may be made the explicit focus of the rating;[15] 4) Ratings may be abandoned in favour of objective measures.

Internal cohesion of the tests Sign-and symptom based items are directed at more element-

Table 1 Correlations (× 100) of items in the Webster Scale with items in the NUDS. Values are also provided for the Total Webster Score, Total NUDS, Total Self Care (SC) and Karnofsky Performance Score

	Northwestern University Disability Scale						*NUDS*	
	Walking	*Dressing*	*Hygiene*	*Eating*	*Feeding*	*Speech*	*Total*	*Karnofsky*
Webster								
Bradykinesia	49	55	50	38	45	42	58	46
Rigidity	25	29	26	28	39	34	35	31
Posture	44	47	48	43	51	39	55	59
Arm swing	40	43	51	38	45	20	48	45
Gait	59	62	59	39	50	28	62	62
Tremor	34	27	47	19	42	6	36	19
Facies	25	36	41	41	43	66	50	25
Seborrhoea	28	43	34	7	42	33	39	32
Speech	17	27	24	35	37	80	43	26
Self care	65	54	63	55	52	42	68	71
Webster Total	65	71	74	57	74	62	82	70
SC Total							81	83

Table 2 Items from the Webster and Self Care Tests Scales with predictive utility for the NUDS

	Regression coefficient	*Standard error*
Webster items		
Bradykinesia	0·285	0·11
Gait	0·131	0·11
Tremor	0·153	0·09
Speech	0·157	0·12
Self care items		
Food preparation	0·248	0·18
Climbing stairs	0·216	0·11
Mobility	0·285	0·18

ary features of movement and its impairment than are ADL items. Performance of ADL items is influenced by a large number of features, several of which are likely to be shared by different items. These considerations suggest that the internal cohesiveness of the items comprising a largely sign-and-symptom based scale like the Webster might be lower than that of an ADL-based instrument. In pursuit of this question we calculated separately for the Webster and the Self Care (both comprising similar numbers of items and using the same four-point scale) the intercorrelations amongst their constituent items. These are shown in table 3. To make the matrices easier to inspect only those values that were highly significant ($p < 0·001$) are displayed. It is clearly evident that the Self Care scale is the more cohesive test. In the Self Care scale, all the correlations that fail to meet our strict significance level involve either eating or tool use, items that involve fine motor control and lack a major component of mobility. In contrast, only a small minority of the Webster inter-correlations met our significance level.

To explore further the relationships between the constituents of the Webster and Self Care scales a principal components analysis[16] was conducted on each. This statistical technique is directed towards the identification of coherent subsets within a group of variables. Subsets of variables (in this case, test items) that are correlated with one another but dissociated from other subsets are combined into "components". The essential feature of the technique is that the original variables are transformed into a new set, the components, which are themselves uncorrelated. These components are ordered so that they account for decreasing proportions of the variance in the original data. Since a small number of these new variables (each of which is a linear combination of the original) may account for a large part of the variance, they may provide a parsimonious summary of the data. In some cases the derived variates, are subjected to a mathematical process known as "rotation" which is intended to aid the identification of the components. The technique may be used descriptively to summarise the data, with redundancy removed. It may also be used interpretatively in an attempt to uncover, for instance, the ability factors that underlie performance.

Table 3 Inter item correlations significant at $p < 0·001$ for the Self Care Scale and the Webster Scale

Self care items		2	3	4	5	6	7	8	9	10	11	12
Dressing	1	0·52	0·60	0·62	0·59	0·63	0·54	0·50	0·48	0·50	0·66	0·60
Eating	2		0·50							0·42		0·40
Cooking	3			0·86	0·50	0·49	0·58	0·58	0·68	0·50	0·78	0·77
House work	4				0·53	0·54	0·50	0·56	0·68		0·79	0·85
Rise from bed	5					0·68	0·63	0·71	0·64	0·50	0·60	0·58
Turn in bed	6						0·60	0·60	0·58	0·43	0·52	0·61
Rise from chair	7							0·57	0·61	0·51	0·65	0·47
Stair climb	8								0·73	0·49	0·68	0·58
Toilet	9										0·73	0·70
Tool use	10										0·53	0·50
Bathing	11											0·77
Mobility	12											

Webster items		2	3	4	5	6	7	8	9	10
Bradykinesia	1	0·54	0·51	0·40						
Rigidity	2		0·52							
Posture	3			0·53	0·51					0·48
Arm swing	4				0·55					0·50
Gait	5									0·54
Tremor	6									
Facies	7								0·74	
Seborrhoea	8									
Speech	9									
Self care	10									

Table 4 Principal component (PC) results for the intercorrelation of the items in the Self Care Scale

	Before rotation		*After rotation*	
	PC1	*PC2*	*PC1*	*PC2*
Dressing	0·77	0·24	0·50	0·54
Eating	0·51	0·73	0·00	0·89
Cooking	0·85	0·22	0·57	0·67
House work	0·84	0·06	0·65	0·53
Rise from bed	0·78	−0·36	0·84	0·15
Turn in bed	0·75	−0·20	0·73	0·27
Rise from chair	0·75	−0·10	0·67	0·35
Stair climb	0·79	−0·33	0·84	0·18
Toilet	0·82	−0·26	0·82	0·26
Tool use	0·65	0·20	0·41	0·54
Bathing	0·88	0·00	0·72	0·51
Mobility	0·85	0·07	0·66	0·55

Tables 4 and 5 show the principal component results obtained before and after a "varimax" rotation, for the Self Care items and Webster items, respectively. For the Self Care items, the two components extracted from the correlations accounted for 70% of the variance. After rotation, the first had loadings above 0·60 on (in descending order) rising from bed, climbing stairs, toilet, turning in bed, bathing, rising from chair, housework and mobility out of doors. Only eating and cooking had such loadings on the second component. These results suggest that the solution for the Self Care scale is rather simple, with one major component that seems to reflect mobility and whole body movements. The minor component seems to reflect manual fine motor coordination.

The three components extracted from the Webster inter-item correlations together only accounted for 65% of the variance. Loading above 0·60 on the first rotated component were arm swing, gait, self care and posture; on the second component, speech and facies; on the third component, seborrhoea. These components are not easy to interpret, although the first appears to relate to mobility. A factor analytic study of a similar set of symptom-based items has been reported by Reynolds and Montgomery.[17] They also required to posit three factors to account for 70% of the variance, with the first factor seeming to be one of mobility.

In summary, these analyses suggest that the Self Care scale is very much more internally

Table 5 Principal component (PC) results for the intercorrelation of items in the Webster Scale

	Before rotation			*After rotation*		
	PC1	*PC2*	*PC3*	*PC1*	*PC2*	*PC3*
Bradykinesia	0·68	0·00	0·35	0·36	0·31	0·60
Rigidity	0·67	0·21	0·21	0·29	0·50	0·44
Posture	0·82	−0·11	−0·01	0·67	0·34	0·34
Arm swing	0·66	−0·49	−0·20	0·84	−0·04	0·14
Gait	0·65	−0·34	−0·33	0·80	0·10	0·01
Tremor	0·37	−0·47	0·36	0·37	−0·25	0·53
Facies	0·60	0·61	−0·10	0·16	0·84	0·09
Seborrhoea	0·27	0·05	0·77	−0·14	0·08	0·80
Speech	0·53	0·73	−0·15	0·07	0·91	0·00
Self care	0·70	−0·14	−0·31	0·72	0·29	0·03

cohesive than the Webster scale and has a simpler underlying structure. Two factors may contribute to these differences between the two scales. First, sign-based test items come closer to reflecting independent, elementary features of the disorder, whereas the ADL items draw upon overlapping sets of elementary features. This in turn implies that whilst a brief ADL-based scale may efficiently summarise the patient's functional status, more detailed investigations into the various features of the disease and their responsiveness to therapy will be better served by a sign-based instrument. Second, it is worth bearing in mind that the apparent cohesiveness of the Self Care scale may partly be due to reliance on patients' self ratings, resulting in some lack of independence in the assessment of items.

General conclusions

1) It should be possible to construct a short, highly cohesive scale, consisting of ADL items, and relying largely on self ratings, which would provide a useful assessment of a patient's general functional status.

2) While ADL-based scales are more internally cohesive than sign-based ones, the relationship between these contrasting types of tests is sufficiently close to provide reassuring evidence of convergent validity.

3) Sign-based scales offer a better prospect than ADL based scales as instruments for analysing patterns of impairment or selective effects of therapy. However, as Study 1 showed, in their present form their reliability is unsatisfactory even when a scale with only four levels is employed.

4) The major source of unreliability seems to be inherent peculiarities and ambiguities in particular performances, rather than expertise of the rater. This might be considerably reduced by careful selection and standardisation of items and their scoring criteria. However, even with reliability optimised, the discriminative power of a test based on subjective clinical ratings is likely to be severely limited. Moreover, greatly extending the number of items will not, in itself, obviate this difficulty. To discriminate between increasingly refined therapeutic interventions, either some means must be found to overcome the limitations of the raters' capacity for absolute categorical judgements or valid objective measures have to be developed.

This research was supported by grants from the Medical Research Council, the Parkinson's Disease Society and a National Advisory Board award. We are grateful to John Gosling and Christian Lueck for helpful comments, to Sandoz Pharmaceuticals for the opportunity to gather the neurologists' ratings in Study 1 and for the test data in Study 2, and to the United Kingdom Parkinson's Disease Research Group for access to their data-base.

1 Marsden CD, Schachter M. Assessment of extrapyramidal disorders. *Brit J Clin Pharmacol* 1981;11:129–51.

2 Potvin PE, Tourtellotte MD. *Quantitative examination of neurological functions*. Boca Raton, Florida: CRC Press, 1984.

3 Webster DD. Clinical analysis of the disability in Parkinson's diseaase. *Mod Treat* 1968;5:257–82.

4 Fahn S, Elton RL. Members of the UPDRS Committee. *Unified Parkinson's disease rating scale*.

5 Siegel S. *Nonparametric statistics for the behavioral sciences*. London: McGraw Hill, 1956.

6 Canter CD, de la Torre A, Mier M. A method for evaluating disability in patients with Parkinson's disease. *J Nerv Ment Dis* 1961;133:143–7.

7 Kennard C, Munro AJ, Park DM. The reliability of clinical assessment in Parkinson's disease. *J Neurol Neurosurg Psychiatry* 1984;47:322–3.

8 Ginanneschi A, Degl'Innocenti F, Magnolfi S, *et al*. Evaluation of Parkinson's disease: reliability of three rating scales. *Neuroepidemiology* 1988;7:38–41.

9 Everitt BS, Hand D, Barrie MA, *et al*. (United Kingdom Bromocriptine Research Group) Bromocriptine in Parkinson's disease: a double-blind study comparing "low-slow" and "high-fast" introductory dosade regimens in de novo patients. *J Neurol Neurosurg Psychiatry* 1989;52: 77–82.

10 Hoehn MM, Yahr MD. Parkinsonism: Onset, progression and mortality. *Neurology (Cleveland)* 1967;5:427–42.

11 Karnofsky DA, Abelmann WH, Craver LF, Burchetval JH. The use of the nitrogen mustards in the palliative treatment of carcinoma, with particular reference to bronchogenic carcinoma. *Cancer* 1948;1:634–9.

12 Brown RG, MacCarthy B, Jahanshahi M, Marsden CD. Accuracy of self-reported disability in patients with Parkinson's disease. *Arch Neurol* 1989;46:955–9.

13 Diamond SG, Markham CH. Evaluating the evaluations: or how to weigh the scales of Parkinsonian disability. *Neurology (Cleveland)* 1983;33:1098–9.

14 Larsen TA, LeWitt PA, Calne DB. Theoretical and practical issues in assessment of deficits and therapy in Parkinsonism. In: Calne DB, ed. *Lisuride and other dopamine agonists*. New York: Raven Press, 1983.

15 Feinstein AR, Josephy BR, Wells CK. Scientific and clinical problems in indexes of functional disability. *Annals of Int Med* 1986;105:413–20.

16 Everitt B, Dunn G. *Advanced methods of data exploration and modeling*. London: Heinemann, 1983.

17 Reynolds NC, Montgomery GR. Factor analysis of Parkinson's impairment. *Arch Neurol* 1987;44:1013–16.