Movement Disorders
Vol. 17, No. 5, 2002, pp. 867–876

Systematic Evaluation of Rating Scales for Impairment and Disability in Parkinson's Disease

Claudia Ramaker, MD,[1]* Johan Marinus, MSc,[1] Anne Margarethe Stiggelbout, PhD,[2] and Bob Johannes van Hilten, PhD, MD[1]

[1]*Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands*
[2]*Department of Medical Decision Making, Leiden University Medical Center, Leiden, The Netherlands*

Abstract: We assessed the clinometric characteristics of rating scales used for the evaluation of motor impairment and disability of patients with Parkinson's disease (PD), conducting a systematic review of PD rating scales published from 1960 to the present. Thirty studies describing clinometrics of 11 rating scales used for PD were identified. Outcome measures included validity (including factor structure), reliability (internal consistency, inter-rater, and intrarater) and responsiveness. We traced three impairment scales (Webster, Columbia University Rating Scale [CURS] and Parkinson's Disease Impairment Scale), four disability scales (Schwab and England, Northwestern University Disability Scale [NUDS], Intermediate Scale for Assessment of PD, and Extensive Disability Scale), and four scales evaluating both impairment and disability (New York University, University of California Los Angeles, Unified Parkinson's Disease Rating Scale [UPDRS], and Short Parkinson Evaluation Scale). The scales showed large differences in the extent of representation of items related to signs considered responsive to dopaminergic treatment or to those signs that appear late in the disease course and lack responsiveness to treatment. Regardless of the scale, there was a conspicuous lack of consistency concerning inter-rater reliability of bradykinesia, tremor, and rigidity. Overall disability items displayed moderate to good inter-rater reliability. The available evidence shows that CURS, NUDS, and UPDRS have moderate to good reliability and validity. In contrast to their widespread clinical use for assessment of impairment and disability in PD, the majority of the rating scales have either not been subjected to an extensive clinometric evaluation or have demonstrated clinometric shortcomings. The CURS, NUDS, and UPDRS are the most evaluated, valid, and reliable scales currently available. © 2002 Movement Disorder Society

Key words: rating scales; Parkinson's disease; systematic review

Parkinson's disease (PD) is a progressive neurological disorder that gradually results in an accumulating disability. Because most of the motor features result from striatal dopamine deficiency, the treatment of patients with PD has focussed on the administration of dopaminergic drugs to alleviate symptoms. New insights in the pathophysiology of PD and an increasing awareness of factors that contribute to levodopa-induced motor complications have stimulated the development of not only new drugs but also very promising surgical techniques.[1–3] Consequently, the increasing number of therapeutic interventions in PD, has highlighted the importance of measuring clinical outcomes. In 1981, Marsden and Schachter[4] reviewed all methods for the assessment of extrapyramidal disorders and presented a comprehensive summary of subjective and objective assessments, regardless of their validity and reliability. Since the appearance of this review the evaluation of patient outcomes, clinometrics, has developed in a science of its own. Information on validity, reliability and responsiveness is now considered as essential knowledge to assure the useful application of a rating scale.[5] We conducted a systematic review of the clinometric aspects of scales that are used by observers to evaluate the motor impairment and disability of patients with PD.

*Correspondence to: Dr. J.J. van Hilten, Department of Neurology, Leiden Universit Medical Center, PO Box 9600, RC Leiden, The Netherlands. E-mail: crag@hetnet.nl

Received 10 September 2001; Revised 13 February 2002; Accepted 12 March 2002

Published online 6 May 2002 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/mds.10248

METHODS

Studies were included if they evaluated clinometric properties of a PD rating scale that addressed impairment

or disability, scored or evaluated by an observer. Self-reporting scales and quality of life measures were therefore excluded from this review.[6] Impairment is defined as an abnormality of a body or organ structure or function; and disability as a reduction of a person's ability to perform a basic task.[7,8] Scales that assessed primarily dyskinesias or motor fluctuations were also excluded.

Search Strategy

The following sources were used to identify studies of interest: Computerized searches of Medline and EMBase using text words (rating) scale, impairment, disability, clinometrics, evaluation, and the individual scale names in combination with "Parkinson" and related terms (search conducted December 2001), reference lists of the reviews found by the Medline and EMBase search-strategy, SCIsearch, the Cochrane Library,[9] symposia reports, PD handbooks, and reference lists of all included publications. Searches were not restricted to the English language.

Methods of Review

Two reviewers independently reviewed the identified publications according to a two-step review process. First, abstracts were reviewed for eligibility. Eligible reports were judged against a set of methodological criteria in which both thoroughness (methodological and statistical) and results of studies testing validity, reliability, and responsiveness were assessed. A checklist was used to evaluate sample characteristics, outcome measures, appropriateness of statistical analysis, and methodological quality. The method of presenting the quality of scales was adopted from McDowell and Newell.[10]

In attempting to interpret the different indices of correlation and degrees of agreement, we noted that there is no general agreement about how high they should be. Because a new rating scale is generally not designed to replicate precisely the existing method against which it is compared, the expected correlation should not be perfect as this may indicate that the new scale is redundant. Few studies, however, declare what levels of correlation are to be taken as demonstrating adequate validity or reliability.

We interpreted the different correlations and degrees of agreement for validity and reliability as follows: The Spearman's coefficient ρ, Pearson's coefficient ρ, Kendall's coefficient W or T, Eta coefficient, and Cramer's coefficient V with values of 0.7 and lower were considered poor,[11] whereas values over 0.7 were considered moderate to good. The values for κ, κ_w and ICC of 0.40 or lower were considered to indicate poor agreement, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial and values over 0.81 good to almost perfect agreement.[12] Cronbach's α lower than 0.70 were considered poor, whereas values of 0.71 to 0.90 were considered moderate to good.[10,13] If, however, α is too high, ($\alpha > 0.90$), then this may reflect redundancy, indicating that some of the items are unnecessary.[11]

The thoroughness of the evidence was classified as follows. If the appropriate statistical procedures were used, the sample size was considered large enough and all circumstances were optimal (i.e., the PD population) then it was classified as good. If less preferable statistical procedures were used or the circumstances were less optimal, then it was classified as substantial. If inappropriate statistical procedures were used or circumstances were less optimal it was classified as moderate, and if the statistical procedure or the circumstances were inadequate, it was classified as poor.

Studies were eligible when they calculated the following clinometric characteristics of disease specific impairment and disability instruments in Parkinson's disease: validity (content validity, criterion validity, and construct validity including factor structure), reliability (internal consistency, inter-rater reliability, intrarater reliability) or responsiveness.

Validity is the extent to which an instrument measures what it is supposed to measure and does not measure what it is not supposed to measure. Three types of validity are frequently discussed: content, criterion, and construct validity.

Content Validity.

Content validity consists of a judgment of whether the instrument samples all the relevant or important contents or domains. It relies on expert opinions and reviews of the literature.

Criterion Validity.

The demonstration of the concordance of an assessment compared with a particular standard, the criterion. It is assessed using correlation coefficients of concordance, or percentage of agreements. The most commonly used correlation coefficients of concordance are Spearman's coefficient ρ, Pearson's coefficient ρ, Kendall's coefficient W and Cramer's coefficient V. Coefficients range from −1 (indicating an inverse linear association) through 0 (indicating no association at all) to +1 (indicating perfect positive linear association). This concept is particularly useful when an obvious gold standard exists for use as a criterion.

Construct Validity.

Construct validity is commonly used instead of criterion validity, because in most cases a gold standard is

lacking. It is demonstrated by examining the relations among a newly created test and other test to show that the new test measures the same construct. Factor analysis is commonly used to study the internal structure of a scale that contains separate components, each reflecting a different aspect of the measured domain. Using this technique a large number of interrelated items are reduced to a smaller number of common dimensions or factors (clusters of items). Unrelated items should not belong to the same factor.

Reliability is the extent to which an instrument is free of measurement error. Reliability assessment aims to quantify the most important sources of measurement error, including both consistency among scale items and reproducibility between and within observers.

Internal Consistency.

Internal consistency estimates the extent to which all items are measuring the same construct. Cronbach's coefficient α, the most frequently used indicator of internal consistency, represents the average of all correlations between all items grouped in all possible combinations of two scale halves. Coefficient α will be equal to zero, when there is no linear relationship between the items. If all items are perfectly reliable and measure the same aspect (true score), then coefficient α is equal to 1. For clinical applications at a patient group level the minimum value is 0.7, for influences at the level of an individual patient, the minimum 0.9 is desirable.[11]

Inter-rater (or Inter-observer) Reliability.

This measures the agreement among different observers performing the assessment on a same individual. Inter-rater reliability is best assessed by the intraclass correlation (=ICC) or the kappa (=κ) statistics.[14] ICC is a parametric measure of agreement and represents the proportion of variance among patients that is caused by true differences.[15] Kappa, developed for the study of nonparametric ratings by observers, measures agreement corrected for the extent of agreement expected by chance alone. Where the categories are ordered, it may be preferable to give different weights to disagreements according to the magnitude of the discrepancy, the κ_w (= weighted kappa).[16] If a squared weighting scheme is used, then the κ_w is identical to the ICC.

Intrarater (or Intra-observer) Reliability.

This measures the reproducibility of the assessment by the same examiner, during repeat assessment (test–retest reliability). The intrarater reliability is also best assessed by the ICC or the κ statistics.

Responsiveness or sensitivity to change is the ability of an instrument to reflect underlying changes over time. In contrast to the assessment of individual differences in change, there is no clear consensus as to how this should be assessed for a rating method.[15,17]

Other information that was gathered included the type of scale, the number of items, the scoring method, and administration time. Whenever information on studies or scales was unclear or incomplete, we contacted the authors with the request to provide additional information.

RESULTS

Description of Studies

Over the period of 1966 to December 2001, 30 studies were identified that described clinometric characteristics of 11 rating scales for patients with PD. We excluded a study by Cutson and colleagues[18] that deals with the Duke University Parkinson's Rating Scale (DUPRS), because the original scale items could not be retrieved. We were unable to trace studies that evaluated responsiveness. Three impairment scales (the Columbia University Rating Scale [CURS], the Parkinson's Disease Rating Scale by Webster [Webster], and the Parkinson's Disease Impairment Scale [PDIS]), four disability scales (the Northwestern University Disability Scale [NUDS], the Intermediate Scale for Assessment of Parkinson's disease [ISAPD], the Schwab and England, and the Extensive Disability Scale [EDS]), and four multimodular scales containing both impairment and disability sections (the New York University Parkinson's disease evaluation [NYU], the University of California Los Angeles scale [UCLA], the Short Parkinson's Evaluation Scale [SPES], and the Unified Parkinson's Disease Rating Scale [UPDRS]) were identified.

We describe clinometric characteristics of individual impairment and disability items. Details on individual scales and a comparison of their clinometric characteristics follow.

Impairment

Content Validity.

In evaluating the content of impairment scales and impairment sections of multimodular scales large differences emerged. Some impairment items were present in all (tremor and bradykinesia) or in the majority (rigidity and gait) of the available scales. Some items were unique for a particular scale (e.g., blepharospasm in the UCLA, short and extra steps in the PDIS). As the core features are not equally represented and defined in the different rating scales, the contribution of these signs to the total score varies from scale to scale (Table 1). The contribution of items dealing with bradykinesia and hypokinesia (including finger and foot taps, successive hand move-

TABLE 1. *Contribution of an item to the total impairment score*

	WEBSTER	UCLA (signs)	CURS	NYU	UPDRS (motor)	PDIS	SPES (motor)
Brady-/hypokinesia	40	23	28	16	37	30	17
Tremor	10	11	20	14	26	20	33
Rigidity	10	9	20	14	19	0	17
Postural stability	0	0	4	0	4	10	8
Other items	40	57	28	56	14	40	25

Values are percentages equal to the possible maximum score for that item/the possible maximum score for the impairment scale or impairment section of multimodular scale.

ments, facial expression, body bradykinesia, akinesia, and arm swing) to the total impairment scores vary from 17% (SPES Motor Evaluation [ME] section) to 40% (Webster). For tremor these values vary from 10% (Webster) to 33% (SPES), for rigidity 0% (PDIS) to 20% (CURS), and for postural stability 0% (Webster, UCLA and NYU) to 10% (PDIS).

Two scales use a weighting factor for each item. In the NYU the maximum possible score for each sign determines the weighting; in the UCLA, as an example, 'akinesia' is weighted nine times whereas mask facies is weighted only once. Several studies repeatedly demonstrated that tremor behaves independent from all other items, not significantly contributing to the explained variance of a scale,[19] nor to the construct validity (Hoehn and Yahr [H&Y] staging).[20,21] Postural instability, an other major feature of PD occurring in the later stages of the disease, is not evaluated in the Webster, the UCLA and the NYU. The item speech is present in five impairment scales or sections (Webster, UCLA, CURS, UPDRS- and SPES ME section). Seborrhea and sialorrhea are evaluated in three (Webster, UCLA, and CURS) and two impairment scales (UCLA and CURS), respectively.

Another problem that emerged concerned the applied methods by which an impairment was evaluated. This was particularly conspicuous for bradykinesia.

Reliability.

Nine studies reported inter-rater reliability of the separate items, whereas only one evaluated intrarater reliability.[22] This study reported a moderate to good intrarater reliability for all items of the CURS, except for rigidity, which was not reported because this study was video-based.

Regardless of the scale, there was a conspicuous lack of consistency among the findings (range, poor to good) concerning inter-rater reliability of the core features bradykinesia, tremor and rigidity as well as for the item speech (Table 2). The majority of the studies found a good inter-rater reliability for postural stability. Seborrhea as well as sialorrhea showed in the CURS a poor[22,23] and in the UCLA a moderate[24] inter-rater reliability.

Disability

Content Validity.

The Schwab and England activities of daily living scale is a staging system, in which 100% stands for completely independent and 0% for a vegetative state. The remaining three disability scales and four disability sections of multimodular rating scales bear only some resemblance in content of items. Dressing, walking, speech, hygiene, and feeding or eating (swallowing) items are included in all scales. Turning in or getting out of bed, and getting out of a chair are included in all scales except in the NUDS. The items handwriting and climbing the stairs are found in four scales (UCLA, NYU, UPDRS Activities of Daily Living [ADL] section and SPES ADL section) and in three (UCLA, EDS, and ISAPD), respectively.

Reliability.

Eight studies reported inter-rater reliability of the separate items, in contrast to the intrarater reliability, which was only evaluated in one study.[20] This study reported a moderate to good intrarater reliability for all items of the PDIS.

Overall, the disability items displayed moderate to good inter-rater reliability, with a few exceptions. Speech scored poor in two studies assessing the NUDS,[23,28] and in one study on the EDS.[28] In the original publication of the UPDRS,[28] Fahn reported a poor inter-rater reliability for walking, in contrast to two later studies that found substantial to excellent values for this item.[21,28]

CLINIMETRIC CHARACTERISTICS OF THE INCLUDED SCALES

Impairment Scales

The three impairment scales (Table 3), the Columbia University Rating Scale (CURS), the Webster, and the Parkinson's Disease Impairment Scale (PDIS), vary in

TABLE 2. *Interpretation of values for interrater reliability*

	Webster	UCLA (signs)	CURS	UPDRS (motor)	SPES (motor)
Brady-/hypokinesia					
Finger tap			$+^{1}$ $++^{22}$	$+^{27}$ $++^{42}$ $+++^{21,26,41}$	$+++^{21}$
Foot tap			$-^{23}$ $++^{22}$	$+^{26,27}$ $+++^{21,41,42}$	
Successive movements	$+^{25}$ $++^{23}$		$-/+^{23}$ $++^{22}$	$-^{26}$ $++^{27,41}$ $++/+++^{21,42}$ $-^{26,42}$	
Facial expression	$-^{23,24}$	$-^{24}$	$+^{22,23}$	$+^{27}$ $++^{21,41}$ $+^{26,27}$	
Body bradykinesia			$-^{23}$ $++^{22}$	$++^{42}$ $+++^{21,41}$	
Akinesia		$+^{24}$			
Arm swing	$-^{23}$ $+^{24}$				
Tremor					
Rest and postural	$++^{23}$ $+^{24,25}$	$+^{24}$	$-/++^{23}$ $+^{25}$ $++^{22}$		
Rest				$+^{27}$ $++/+++^{21,41}$ $+++^{26,42}$	$+++^{21}$
Postural				$++^{21}$ $-^{41}$	$++/+++^{21}$
Action				$+^{21,27}$ $++^{26,42}$ $+^{26,42}$	
Rigidity	$-^{23}$ $+^{24,25}$	$++^{24}$	$-^{25}$ $-/+^{23}$	$++/+++^{21}$ $+++^{27}$	$+++^{21}$
Postural stability			$+++^{22}$	$+^{26,42}$ $+++^{43,27,41}$ $-^{26}$	$+++^{21}$
Posture	$-^{23}$ $+^{24,25}$	$+^{24}$	$++^{22}$	$+^{27,42}$ $++^{21}$ $+++^{41}$ $-^{42}$	
Speech	$-^{23}$ $+^{24}$	$+^{24}$	$-^{23}$ $+^{22}$	$+^{27}$ $++^{21,26}$ $+++^{41}$	$++^{21}$
Seborrhoea	$-^{23}$	$+^{24}$	$-^{22,23}$		
Sialorrhoea	$+^{24}$	$+^{24}$	$-^{23}$		

The superscript number corresponds with the studies in References in which interrater reliability per item is evaluated. For the NYU and the PDIS, no information on interrater reliability (per item) is available.

–, poor; +, moderate; ++, substantial; +++, good.

number of items (10, 27, and 10 items respectively) and in scoring of items (0–4, 0–3, and 0–3).

Parkinson's Disease Rating Scale by Webster.

For a scale that has been used for a long time by many investigators, surprisingly little evidence is published on its validity and reliability. Notably, the Webster includes one disability (self-care) and nine impairment items, which makes this scale conceptually unclear. From a factor analysis, assessed in one study, three factors were derived, including (I) arm swing, gait, self-care and pos-

TABLE 3. *Results of validity and reliability and thoroughness (strength of evidence) of validity and reliability testing*

Scale	Scale type[a]	N (items)	Validity		Reliability			No. of studies[c]
			Construct	Factor[b]	Interrater	Intrarater	Internal	
CURS 1969		25	++(+)/+++	/+++	++/+++	+++/+++	+++/+++	5[22,23,25,29,30]
CURS-modified (Sydney) 1993	I	11	++(+)/+++	0	+++/+++	+++/+++	0	1[22]
CURS-modified 1985		8	0	/–	+/+	0	0	1[52]
EDS 1991	D	21	+++/+++	0	+++/+++	0	0	1[25]
ISAPD 1987	I,D	13	+++/+++	/+++	++(+)/+++	0	+++/+++	1[31]
NUDS 1980	D	6	++(+)/+++	0	++(+)/+++	0	0	6[2,19,23–25,28]
NYU 1980	I,D	6	+++/+++	0	0	0	0	1[36]
PDIS 1987	I	10	–(+)/+	/–	0	++(+)/++	0	1[20]
SPES 1997	I,D	25	+++/+++	/+++	+++/++(+)	0	0	1[21]
UCLA 1981	I,D	21	0	0	++(+)/+++	0	0	2[24,28]
UPDRS 1987		31	+++/+++	/+++	++/+++	0	+++/+++	4[26,27,40,42]
UDRS ADL	I,D	13	+++/+++	/+++	0	0	+++/+++	2[21,39]
UPDRS ME		14	+(+)/+(++)	/+++	++/++	0	+++/+++	6[21,32,38,39,41]
Webster 1968	I	10	++/+	/++	–(+)/+++	0	0	6[19,23,24,28,29,51]

Signs before the slash refer to results of validity and reliability and signs behind the slash refer to thoroughness (strength of evidence) of validity and reliability testing. Results of validity and reliability testing: 0, no numerical results reported; ?, results not interpretable; –, poor results; +, moderate results; ++, substantial results; +++, good results.

Thoroughness of validity and reliability testing: 0, no reported evidence; ?, results not interpretable; –, poor evidence; +, moderate evidence; ++, substantial evidence; +++, good evidence.

[a]I, impairment scale; D, disability scale.

[b]Thoroughness of testing only.

[c]Superscript numbers correspond with the studies in References.

ture; (II) speech and facies; (III) seborrhea.[19] Four studies showed that the scale displays poor to moderate interrater reliability.[23–28]

Columbia University Rating Scale.

Although the Columbia University Rating Scale (CURS) has been used frequently in clinical studies before the introduction of the UPDRS in 1981, few studies have been published on the validity and/or reliability of this scale, mostly in combination with other PD rating scales.[22,23,28,29] The available evidence shows the CURS to have moderate to good validity and reliability. The factor structure was evaluated in only one study, which included 95 patients with PD plus syndromes, and thus precludes a conclusion on this issue in PD.[30] A modified version of the CURS, the Sydney scale, appears to be equally valid and reliable.[22]

Parkinson's Disease Impairment Scale.

Only one study has assessed validity and reliability of the Parkinson's Disease Impairment Scale (PDIS). Due to unclear factor analysis and the subsequent assessment of the construct validity based on these factors, the validity of this scale is questionable.[20] The intrarater reliability appeared to be moderate to good.

Disability Scales

Four disability scales, including the Northwestern University Disability Scale (NUDS), the Intermediate Scale for Assessment of Parkinson's disease (ISAPD), the Schwab and England and the Extensive Disability Scale (EDS) are hard to compare, because they vary much in scoring, grading, number, and kind of items. Although the ISAPD is, among others, based on the NUDS, its grading is different; 0 to 3 instead of 0 to 10.

Schwab and England.

The Schwab and England scale has become a standard assessment tool in PD and has been used in hundreds of studies. The clinometric properties of this scale, however, have never been established. The data available from studies with a primary aim to investigate characteristics of other rating scales suggest a moderate to substantial validity and good reliability.[28,31,32]

Northwestern University Disability Scale.

Two studies found a moderate to good construct validity.[19,28] These studies showed that the total Northwestern University Disability Scale (NUDS) score correlates highly with the total Webster score (Kendall's W = 0.82)[19] and with the CURS (Spearman's ρ = –0.78),[28] which are both impairment scales. The interrater reliability of the NUDS was found to be excellent by its designers[33] but only moderate by others.[23,24,28] A reason for the latter could be the combined effect of the large number of severity gradations in this scale and the use of non-weighted κs. Although this scale is frequently used, no information is available on internal consistency or intrarater reliability.

Intermediate Scale for Assessment of Parkinson's Disease.

Evaluated only by its designers, the Intermediate Scale for Assessment of Parkinson's disease (ISAPD) shows a moderate to good correlation with the H&Y, with the UPDRS and with the Schwab and England.[31] In the same study, the results were also excellent for the internal consistency and good for the inter-rater reliability. The administration time was recorded as 7 minutes (±3.70).[31]

Extensive Disability Scale.

The Extensive Disability Scale (EDS) is a modified version of the Minimal Record of Disability (MRD),[34,35] which is used in examining patients suffering from multiple sclerosis and has only been used and tested by its authors, who found a moderate to good construct validity and inter-rater reliability.[28] The administration time was stated as 15–20 minutes by a trained reviewer.[28]

Impairment and Disability Sections in Multimodular Scales

In comparing the four impairment and disability scales, the New York University Parkinson's disease evaluation (NYU), the Short Parkinson's Evaluation Scale (SPES), the University of California Los Angeles scale (UCLA), and the Unified Parkinson's Disease Rating Scale (UPDRS), we noticed the similarity in item content. All scales included items such as bradykinesia, tremor, rigidity, walking, eating, turning in bed, and handwriting.

New York University Parkison's Disease Evaluation.

For this scale only poor construct validity with the H&Y was reported.[36] The administration time was stated as 10 minutes by a trained examiner.[36]

University of California Los Angeles Scale.

The UCLA scale is rarely used in clinical trials and beyond the work of Martínez-Martín,[24] who found a moderate to good inter-rater reliability, no further evidence for reliability or validity of the scale has been published.

Unified Parkinson's Disease Rating Scale.

The UPDRS has found broad acceptance for the evaluation of PD and has been used in many trials.[37] Nine studies extensively tested and evaluated this scale. Like the Webster, the UPDRS ADL section is conceptually unclear as it includes several impairment items (salivation, falling, freezing, tremor, and sensory complains). Nevertheless, the UPDRS demonstrates high internal consistency and inter-rater reliability, shows moderate construct validity, and has a stable factor structure.[21,28,32,38–42] Even across *off*- and *on*-state examinations, the ME section of this scale has a stable factor structure and high internal consistency.[32] The high internal consistency of the ADL and motor section most likely indicates a redundancy of items. This was underscored by a previous study that successfully reduced the ADL and motor section of the UPDRS to eight items each, without losing reliability or validity.[39] The time to administer was stated 10–20 minutes[28] and assessed as 16.95 minutes (±7.98).[28]

SPES.

Evidence for construct validity and inter-rater reliability of the SPES is good, but was only reported in an article by its original designers.[21] The advantage of the SPES seems to be that it is short, and easy to administer in 7–10 minutes (by neurologists).[21]

DISCUSSION

Compared to their widespread clinical use for assessment of impairment and disability in PD, rating-scales are seldom extensively evaluated for validity and reliability. The terms impairment and disability are derived from the World Health Organization International Classification of Impairments, Disabilities, and Handicaps (ICIDH; http://www.who.int/icidh).[7,8] The ICIDH-2 was developed recently, and introduces new terms; body function and structures are handled both positive (functional and structural integrity) and negative (impairment) as well as activities (activity vs. activity limitation).

Systematically reviewing the available literature, we traced 30 studies describing clinometric issues of 11 scales for impairment and disability rating in PD. In general, a criticism could be made on the frequent choice of the H&Y as the gold standard for testing other scales, because, to the best of our knowledge, none have evaluated its clinometric data. Nevertheless, the H&Y is the most commonly used method of establishing the severity of PD with a simple staging assessment.

In evaluating impairment items, the contribution of the core motor features of PD to the total impairment score appears to vary from scale to scale. For instance, items dealing with brady kinesia and hypokinesia contribute almost 40% to the total score of the UPDRS ME section resulting in a strong effect on the sum scores of the impairment section and on the total score.

There are also large differences in the extent of representation of items related to symptoms considered responsive to dopaminergic treatment (e.g., bradykinesia, rigidity) or those that appear late in the disease course and lack responsiveness to dopaminergic treatment (e.g., postural instability, swallowing, speech, freezing).

Hence, these differences in content should be taken into consideration when choosing a scale for evaluating a short-term dopaminergic treatment or a long-term follow-up in which the occurrence of signs not responsive to dopaminergic treatment indicate disease progression. Generally, within the framework of impairments, items as sialorrhea and seborrhea have a limited clinical significance. Regardless of the scale, the findings concerning inter-rater reliability of the core features bradykinesia, tremor and rigidity as well as for the item speech lacked consistency. The majority of the studies, however, found a good inter-rater reliability for postural stability. Clearer description of items may help to improve inter-rater reliability of items. To avoid the problems with inter-rater reliability, objective measurements could be considered in assessing impairment in PD.[43–48] It is remarkable that only one study evaluated intrarater reliability on this level of disease assessment, which is relevant in the case of longitudinal studies performed by one assessor.

Although there is general agreement on the definition of disability (i.e., the experienced difficulty in carrying out activities of daily living), there is no consensus on what should be measured. All evaluated disability scales and sections included the items of the NUDS (dressing, walking, speech, hygiene, feeding, and eating). Overall disability items displayed moderate to good inter-rater reliability. The low inter-rater reliability values repeatedly found for speech and walking suggest that these items are difficult to score or lack clear anchors.

The PD rating scales identified can be divided in three groups: impairment scales, disability scales, and multimodular scales containing both impairment and disability sections. By comparing the three impairment scales Webster, CURS, and PDIS, we found evidence for the CURS to have strong validity, where there is insufficient data on validity available for the Webster and the PDIS. As the overall reliability of the CURS is moderate to good, the inter-rater reliability of the Webster is assessed as poor to moderate. So, as a brief rating method the Webster appears adequate, but the available clinometric data on CURS point out that this scale is preferable. The PDIS has inadequately been evaluated by its designers and due to the lack of other information on clinometric issues of the PDIS, no recommendations can be given with respect to this scale. The four disability scales, the NUDS, the ISAPD, the Schwab and England and the EDS bear hardly any resemblance. Large differences between the scales are found in the scoring and grading of items. The Schwab and England disability scale takes a unique position, because this scale uses a different grading system and has never been primarily evaluated for its clinometric characteristics. The construct validity and the inter-rater reliability of the NUDS, ISAPD and EDS were found to be moderate to good, suggesting no preference. Only the NUDS was evaluated independently. The ISAPD, evaluated only by its designers, appears to be a very valid and reliable disability scale, which may be useful as a tool for evaluation of disability in PD. Independent verification of the clinometric characteristics, however, is recommended.

Of the scales containing both an impairment and a disability section, the UPDRS is the most widely used and tested scale. The NYU, SPES, and UCLA are rarely used and have only been evaluated by the designers. The construct validity of the UPDRS is satisfactory in those studies that have used the H&Y as comparison. Important differences between these scales include the scoring and the contribution of the individual items to the subtotal and total score. In relation to the validity aspects of the UPDRS, some findings deserve comments. The construct validity of the UPDRS has to be considered very satisfactory. The UPDRS ADL section, however, is conceptually unsound as it includes several impairment items. Concerning the inter-rater reliability, the UPDRS, the SPES, and the UCLA should be considered reliable scales. The SPES and UCLA, however, were evaluated only by designers of the scales. The UPDRS demonstrates a very high internal consistency, but the effects of redundancy (several items focused on the same aspect of the construct) should be kept in mind. Internal consistency increases with the number of items and depends substantially on the homogeneity of the items and on the inter item correlation. Taken together, the evaluation of the impairment and disability sections as a whole show that the UPDRS is a reliable and valid scale, although these sections include some redundant and unreliable items. The SPES appears to be a valid and reliable scale that might be considered for evaluation of patients with PD. Nonetheless, independent verification of the clinometric characteristics is recommended. Because the UCLA and NYU lack thorough clinometric testing, no recommendations can be given.

Others have reviewed disease-specific PD scales,[4,43,49,50] but only Mitchell and associates[37] presented some clinometric properties of the most commonly used scales (identified through a Medline search conducted from 1966 until August 1998). In this study the UPDRS was found to be the most thoroughly studied scale with overall better clinometric properties compared to other scales. As mentioned by the authors, one of the limitations of this study lies in the main focus, which was not to summarize the clinometrics of scales but to examine the pattern of utilization of disease-specific clinical scales used

as endpoints in PD trials. The summary of clinical properties they report is simple and is intended to serve as a guide.

In summary, this review underscores that the clinometric soundness of the majority of PD assessment scales is questionable. Moreover, as these scales are generally used in trials on PD patients who lack serious comorbidity, there is no information on the clinometric behavior of the scales in unselected PD populations.

We emphasize the following critical notes regarding clinometric issues:

1. The most important question in choosing a scale is how well it is suited to the task at hand in terms of validity, reliability, and efficiency.
2. A greater number of items increases the internal consistency and leads to greater concordance between examiners (its reliability increases). Reliability of a composite scale will increase as a function of the number of the individual items that are included. Limiting the number of items in a scale, however, contributes to simplicity and utility of the assessment, at the expense of completeness, sensitivity, and reliability.
3. It is remarkable that none of the studies addressed differences in responsiveness between scales, which is required to ensure the usefulness in the longitudinal evaluation of PD. Responsiveness is an essential part of the statistical analysis as it refers to the ability of a measure to reflect change.
4. Video recordings may help to improve assessment of inter- and intrarater reliability in studies. These recordings have their limitations, however, for they can only be used to score items that are clearly visible or audible. Rigidity, seborrhea and sialorrhea are difficult to discern on tape and should not be included if a scale is used for video assessments.

Acknowledgments: We thank Dr. F. Durif, Dr. M. Hely, Prof. L. Henderson, Prof. Dr. C. Kennard, Dr. P. Martínez-Martín, Prof. Dr. J. Opara, Dr. J. M. Rabey, and Dr. N.C. Reynolds, Jr. for providing additional information. C.R. is funded by the Prinses Beatrix Fonds (project no. 97-0205) and J.M. is funded by the Netherlands' National Research Council (project no. 440-33-02).

REFERENCES

1. Rascol O, Brooks DJ, Korczyn AD, De Deyn PP, Clarke CE, Lang AE. A 5-year study of the incidence of dyskinesia in patients with early Parkinson's disease who were treated with ropinirole or levodopa. N Engl J Med 2000;342:1484–1491.
2. Parkinson Study Group. Pramipexole vs. levodopa as initial treatment for Parkinson disease: a randomized controlled trial. JAMA 2000;284:1931–1938.
3. Lang AE. Surgery for levodopa-induced dyskinesias. Ann Neurol 2000;47(Suppl.):S193–S199.
4. Marsden CD, Schachter M. Assessment of extrapyramidal disorders. Br J Clin Pharmacol 1981;11:129–151.
5. Handbook of neurological rating scales. New York: Demos Vermande; 1997.
6. Marinus J, Ramaker C, van Hilten JJB, Stiggelbout AM. Health related quality of life in Parkinson's disease: a systematic review of disease specific instruments. J Neurol Neurosurg Psychiatry 2002; 72:241–248.
7. World Health Organization. International classification of impairments, disabilities, and handicaps : a manual of classification relating to the consequences of disease. Geneva: World Health Organization, 1980.
8. Simeonsson RJ, Lollar D, Hollowell J, Adams M. Revision of the International classification of impairments, disabilities, and handicaps: developmental issues [see comments]. J Clin Epidemiol 2100;53:113–128.
9. The Cochrane Controlled Trials Register. The Cochrane Library Issue 3. 2001. Oxford, UK: Update Software; http://www.updatesoftware.com/cochrane
10. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires, Second ed. New York: Oxford University Press; 1996.
11. Nunnally JC. Psychometric theory. Third ed. New York: McGraw-Hill; 1994.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
13. Feinstein AR. Clinometrics. First ed. New Haven: Yale University Press; 1987.
14. Fleiss JL. The measurement of inter-rater agreement. In: Fleiss JL, editor. Statistical methods for rates and proportions. New York: John Wiley; 1981. p 212–236.
15. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Second ed. New York: Oxford University Press; 1995.
16. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70: 233–220.
17. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459–468.
18. Cutson TM, Sloane R, Schenkman M. Development of a clinical rating scale for persons with Parkinson's disease. J Am Geriatr Soc 1999;47:763–764.
19. Henderson L, Kennard C, Crawford TJ, et al. Scales for rating motor impairment in Parkinson's disease: studies of reliability and convergent validity. J Neurol Neurosurg Psychiatry 1991;54:18–24.
20. Reynolds NC, Jr., Montgomery GK. Factor analysis of Parkinson's impairment: an evaluation of the final common pathway. Arch Neurol 1987;44:1013–1016.
21. Rabey JM, Bass H, Bonuccelli U, et al. Evaluation of the Short Parkinson's Evaluation Scale: a new friendly scale for the evaluation of Parkinson's disease in clinical drug trials. Clin Neuropharmacol 1997;20:322–337.
22. Hely MA, Wilson A, Williamson PM, O'Sullivan DJ, Rail D, Morris JGL. Reliability of the Columbia Scale for assessing signs of Parkinson's disease. Mov Disord 1993;8:466–472.
23. Geminiani G, Cesana BM, Tamma F, et al. Interobserver reliability between neurologists in training of Parkinson's disease rating scales—a multicenter study. Mov Disord 1991;6:330–335.
24. Martínez-Martín P, Carrasco de la Pena JL, Ramo C, Antiguedad AR, Bermejo F. [Study of inter-observer reliability in the use of qualitative scales assessing Parkinson's disease (II)]. Arch Neurobiol (Madr) 1988;51:287–291.
25. Ginanneschi A, Degl'Innocenti F, Maurello MT, Magnolfi S, Marini P, Amaducci L. Evaluation of Parkinson's disease: a new approach to disability. Neuroepidemiology 1991;10:282–287.

26. Fahn S, Elton RL. Unified Parkinson's disease rating scale. In: Fahn S, Goldstein M, Marsden D, Calne DB, editors. Recent developments in Parkinson's disease, Volume II. New Jersey: MacMillan; 1987. p 153–163.
27. Martínez-Martín P, Gil-Nagel A, Morlán Gracia L, Balseiro Gómez J, Martínez-Sarriés FJ, Bermejo F. Unified Parkinson's disease rating scale characteristic and structure. Mov Disord 1994; 9:76–83.
28. Martínez-Martín P, Carrasco de la Pena JL, Ramo C, Antiguedad AR, Bermejo F. [Inter-observer reproducibility of qualitative scales in Parkinson's disease (I)]. Arch Neurobiol (Madr) 1987; 50:309–314.
29. Ginanneschi A, Degl'Innocenti F, Magnolfi S, et al. Evaluation of Parkinson's disease: Reliability of three rating scales. Neuroepidemiology 1988;7:38–41.
30. Baas H, Stecker K, Fischer PA. Value and appropriate use of rating scales and apparative measurement in quantification of disability in Parkinson's disease. J Neural Transm Park Dis Dement Sect 1993; 5:45–61.
31. Martínez-Martín P, Gil-Nagel A, Morlán Gracia L, et al. Intermediate scale for assessment of Parkinson's disease. Characteristic and structure. Parkinsonism Rel Disord 1995;1:97–102.
32. Stebbins GT, Goetz CG. Factor structure of the Unified Parkinson's Disease Rating Scale: motor examination section. Mov Disord 1998;13:633–636.
33. Canter CJ, de la Torre R, Mier M. A method of evaluating disability in patients with Parkinson's disease. J Nerv Ment Dis 1961; 133:143–147.
34. LaRocca MG, Scheinberg LC, Slater RJ, et al. Field testing of a minimal record of disability in multiple sclerosis: the United States and Canada. Acta Neurol Scand Suppl 1984;101:126–138.
35. Slater RJ, LaRocca NG, Scheinberg LC. Development and testing of a minimal record of disability in multiple sclerosis. Ann NY Acad Sci 1984;436:453–468.
36. Lieberman A, Dziatolowki M, Gopinathan G, Kopersmith M, Neophytides A, Korein J. Evaluation of Parkinson's disease. In: Goldstein M, editor. Ergot compounds and brain function: neuroendocrine and neuropsychiatric aspects. New York: Raven Press; 1980. p 277–286.
37. Mitchell SL, Harper DW, Lau A, Bhalla R. Patterns of outcome measurement in Parkinson's disease clinical trials. Neuroepidemiology 2000;19:100–108.
38. Stebbins GT, Goetz CG, Lang AE, Cubo E. Factor analysis of the motor section of the unified Parkinson's disease rating scale during the off-state. Mov Disord 1999;14:585–589.
39. van Hilten JJ, van der Zwan AD, Zwinderman AH, Roos RA. Rating impairment and disability in Parkinson's disease: evaluation of the Unified Parkinson's Disease Rating Scale. Mov Disord 1994;9:84–88.
40. Nouzeilles MI, Merello M. Correlation between results of motor section of UPDRS and Webster scale. Mov Disord 1997;12:613.
41. Goetz CG, Stebbins GT, Chmura TA, Fahn S, Klawans HL, Marsden CD. Teaching tape for the motor section of the unified Parkinson's disease rating scale. Mov Disord 1995;10:263–266.
42. Richards M, Marder K, Cote L, Mayeux R. Inter-rater reliability of the unified Parkinson's Disease Rating Scale for motor examination. Mov Disord 1994;9:89–91.
43. Teravainen H, Calne D. Quantitative assessment of parkinsonian deficit. In: Rinne UK, Linger M, Stamm G, editors. Parkinson's disease: current progress, problems, and management. New York: Elsevier/North-Holland Biomedical Press; 1980.
44. Potvin AR, Tourtellotte WW, Syndulko K, Potvin J. Quantitative methods in assessment of neurological function. CRC Crit Rev Bioeng 1981;6:177–224.
45. Jankovic J. Pathophysiology and clinical assessment of motor symptoms in Parkinson's disease. In: Koller WC, editor. Handbook of Parkinson's disease. New York: Marcel Dekker; 1992. p 99–126.
46. Ringendahl H. [Standardization of a motor performance series for measuring fine motor disorders in Parkinson disease]. Nervenarzt 1998;69:507–515.
47. Lauk M, Chow CC, Lipsitz LA, Mitchell SL, Collins JJ. Assessing muscle stiffness from quiet stance in Parkinson's disease. Muscle Nerve 1999;22:635–639.
48. Caligiuri MP, Galasko DR. Quantifying drug-induced changes in Parkinsonian rigidity using an instrumental measure of activated stiffness. Clin Neuropharmacol 1992;15:1–12.
49. Lang AE, Fahn S. Assessment of Parkinson's disease. In: Munsat TL, editor. Quantification of neurological deficit. Boston: Butterworths; 1989. p. 285–309.
50. Martínez-Martín P. Rating scales in Parkinson's disease. In: Jankovic J, Tolosa E, editors. Parkinson's disease and movement disorders. Baltimore: Williams and Wilkins; 1993. p 281–292.
51. Kennard C, Munro AJ, Park DM. The reliability of clinical assessment of Parkinson's disease. J Neurol Neurosurg Psychiatry 1984; 47:322–323.