Measuring Activities of Daily Living in Parkinson's disease: On a road to nowhere and back again?

Peter Hagell [1, 2]

[1] The PRO-CARE Group, Faculty of Health Science, Kristianstad University, Kristianstad, Sweden

[2] Restorative Parkinson Unit, Division of Neurology, Depratment of Clinical Sciences Lund, Lund University, Lund, Sweden

As accepted for publication in *Measurement* on September 17, 2018.

Reference to final published version:
Hagell P. Measuring activities of daily living in Parkinson's disease: On a road to nowhere and back again? *Measurement* 2019; 132: 109-124.
(doi: 10.1016/j.measurement.2018.09.050)

Abbreviations:

ADL: Activities of daily living
CTT: Classical test theory
DIF: Differential item functioning
ICC: Item characteristic curve
ICF: International Classification of Functioning, disability and health
IRT: Item response theory
MDS-UPDRS: Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale
MID: Minimally important difference
MRFA: Minimum rank factor analysis
PDAQ: Penn Parkinson's Daily Activities Questionnaire
PCA: Principal component analysis
PCM: Partial credit model
PD: Parkinson's Disease
PDQ-39: The 39-item Parkinson's Disease Questionnaire
PRM: Polytomous Rasch model
RCT: Randomized controlled trial
RMT: Rasch measurement theory
RSM: Rating scale model
SEM: Standard error of measurement
UPDRS: Unified Parkinson's Disease Rating Scale

Corresponding author:
Peter Hagell, RN PhD
Faculty of Health Science
Kristianstad University
SE-291 88 Kristianstad
Sweden
Tel: +46 44 250 3956
E-mail: Peter.Hagell@hkr.se

ABSTRACT
Parkinson’s disease (PD) is a progressive neurodegenerative disorder associated with increasing disability and limitations in performance of activities of daily living (ADL) despite availability of effective symptomatic therapy. Following an overview of classical test theory (CTT) and Rasch measurement theory (RMT), the case of a clinical PD trial aiming to demonstrate ADL improvements by using the ADL section (part II) of the Unified PD Rating Scale (UPDRS) to measure ADL outcomes is considered and central questions related to its validity and interpretation are addressed. It is found that while CTT did not detect any issues, RMT in combination with conceptual considerations seriously challenged the role of the UPDRS II as an ADL outcome measure. Results are discussed from historical, methodological and clinical perspectives.

Key words:
Activities of daily living; Classical test theory; Outcome measurement, Parkinson’s disease, Rasch measurement theory; Rating scales

1. Introduction

1.1. Parkinson´s disease

Parkinson´s disease (PD) was first described by English physician James Parkinson in 1817 [1]. PD is a progressive neurodegenerative disorder that affects an estimated 0.3% of the population at large and 1% of people above 60 years of age [2]. The most typical and cardinal features of PD are neurological motor symptoms of bradykinesia (slowness of movement), muscle rigidity, tremor, and postural impairments. However, non-motor features (e.g., depression, anxiety, sleep disorders, fatigue, cognitive impairment, dysautonomia and pain) are also common and make significant contributions to the overall impact of PD [3, 4]. The core pathology believed to cause the main motor symptoms is a striatal dopamine deficit due to progressive loss of nigrostriatal dopaminergic neurons [3]. Symptomatic dopaminergic pharmacotherapy is initially successful but a fluctuating drug response and dyskinesias often develop over time. With the occurrence and progression of both motor and non-motor symptoms, often in complex and fluctuating patterns, the disease is associated with significant consequences in terms of, e.g., deteriorating quality of life and ability to perform activities of daily living (ADL) [5-7].

Although the cause of PD remains enigmatic, there have been considerable advances in our understanding of PD over the 200 years since the disorder first was described [8]. Similarly, therapeutic advances have been significant, particularly during the past 50 to 60 years [8, 9]. Up until the 1960s medical therapy was largely limited to anticholinergic drugs that only offered limited symptomatic relief. By the late 1960s, levodopa revolutionized PD therapy by offering dramatic symptomatic relief. During the following decades, additional therapeutic approaches such as dopamine agonists, enzyme inhibitors and functional neurosurgery have been introduced that together enable long-term symptom control. However, all available therapies hitherto are symptomatic and disease modifying therapies are still lacking.

1.2. Outcome measures in Parkinson's disease

The development of effective symptomatic PD therapy highlighted the need for useful and relevant outcome measures to determine the effectiveness and value of treatments. As in other areas of clinical medicine, such outcome measures have typically consisted of single- or multi-item instruments (or "scales") with defined response categories of various sorts [10]. The perhaps earliest instrument proposed to assess therapeutic benefit in PD was that by Duff in an early clinical trial [11]. This scale included ten "modal activities" of daily life with various degrees of movement complexities (Turning in bed, rising from and returning to bed; Dressing/undressing; Performance of the toilet, especially shaving in men; Eating; Walking; Turning; Climbing stairs; Speaking; Writing; Facial expression); each item was rated by assigning one of six ordered response categories scored 0-5 (No activity; Grossly restricted; Moderate restriction, especially by rigidity; Moderate restriction, especially by tremor; Activity approaching normal, but slow/clumsy; Activity practically normal). Other early PD rating scales include the Schwab & England activities of daily living scale [12] and the Northwestern University Disability Scale [13]. Similar to the Duff scale, these also focused on activity performance rather than symptom severity. However, it was not until following demonstration of the clinical effectiveness of levodopa [14] that instruments for outcome assessment in PD started to proliferate. These included, e.g., the Webster, King's College Hospital, Columbia University, and New York University scales [15]. Although some of these included aspects of activity performance, their primary focus tended to be on assessing the severity of various motor symptoms through standardized neurological examinations. This conceptual shift from what today is categorized as Activities to Body functions [16], was probably due to an intention to evaluate the symptomatic effects of levodopa and other emerging PD therapies.

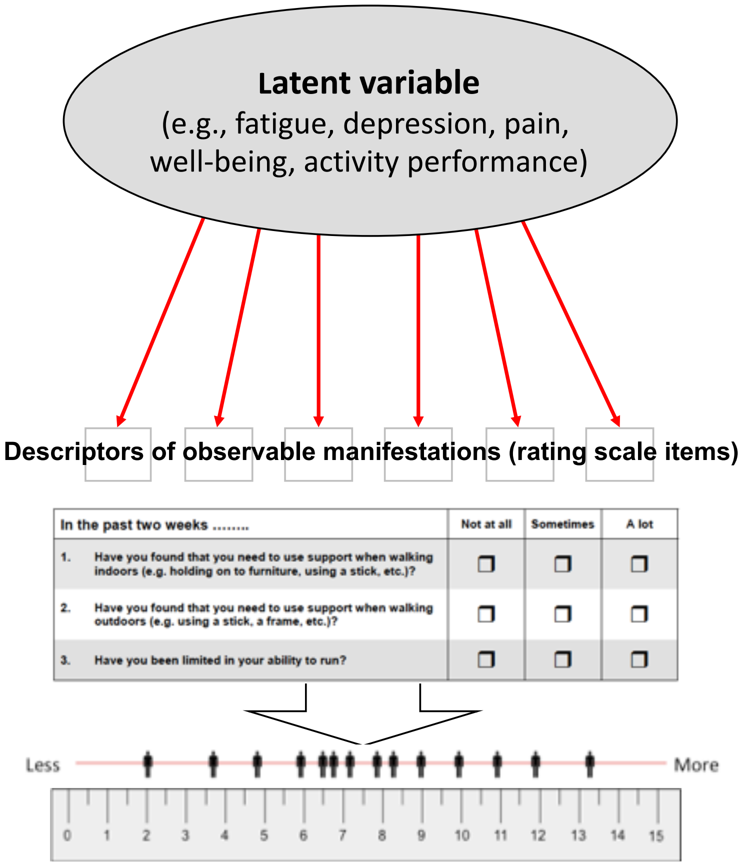
The development and use of similar albeit different instruments by various investigators hampered the possibility to compare results from various studies [17]. This led to the development of the Unified PD Rating Scale (UPDRS), which was based on previously available scales and aimed to overcome the problem of incomparability of study results by introducing a common means of evaluating PD and therapeutic responses [15]. The UPDRS consists of four main parts intended to cover major aspects of PD: Mentation, behavior and mood (part I), Activities of daily living (part II), Motor examination (part III), and Complications of therapy (part IV). More recently, a modification of the UPDRS was conducted by the International Parkinson and Movement Disorder Society (MDS), named the MDS-UPDRS [18, 19]. The basic structure remains intact compared to the original UPDRS but parts I and II were renamed as "Non-motor aspects of experiences of daily living" and "Motor aspects of experiences of daily living", respectively.

2. Psychometrics and rating scales as health outcome measures

At this point, it may be appropriate to briefly review approaches to the development and quality assurance of scales used in the health sciences, as well as in, e.g., education and psychology.

2.1. Basic principles

Most phenomena (variables) of interest in health outcome measurement (e.g., disease severity, quality of life, disability) are not directly observable. Since such latent variables cannot be observed or measured directly, one has to rely on observable manifestations or consequences of the latent variable. These manifestations are operationalized as items (questions, statements, observed behaviour or performance, etc.) that make up the instrument. Variations in the observable manifestations (item responses) are assumed to reflect variations in the latent variable. This principle is illustrated in Figure 1.


Latent variable
(e.g., fatigue, depression, pain, well-being, activity performance)
Descriptors of observable manifestations (rating scale items)
Less
More
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

In the past two weeks	Not at all	Sometimes	A lot
1. Have you found that you need to use support when walking indoors (e.g. holding on to furniture, using a stick, etc.)?	□	□	□
2. Have you found that you need to use support when walking outdoors (e.g. using a stick, a frame, etc.)?	□	□	□
3. Have you been limited in your ability to run?	□	□	□

Fig. 1. Illustration of the basic instrument design and assumptions underpinning latent variable measurement. Items are observable manifestations of the unobservable latent target variable and are expected to reflect variations in the latent variable. Observed item responses form the basis in the measurement process used to locate the individual on a latent quantitative continuum intended to represent her/his position on the target variable, from less to more.

For example, consider that we want to measure activity performance. To do so, we may ask a person to report his or her ability to perform various activities. Based on those responses, we may be able to assess the person's location on a latent activity performance continuum. With reference to Figure 1, based on some conceptual definition of activity performance, items that express relevant activities are presented to the individual, who is asked to indicate her or his performance level on each item according to one of two or more ordered response categories that each describe a certain performance level. To quantify the qualitative information achieved in this manner, each descriptive response category (e.g., none – mild – moderate – severe) is assigned an integral numeral (e.g., 0 – 1 – 2 – 3) as a means of partitioning the underlying latent continuum into successively increasing (or decreasing) amounts of the variable. Item responses are then typically summed into a total score intended as the basis of locating the respondent on a continuum from less to more, in order to describe the level of individuals or groups of people on the latent variable, make comparisons and evaluate changes following therapeutic interventions.

2.2. Classical test theory

The roots of the principles reviewed in section 2.1. can be traced to the late 19th and early 20th centuries, when behavioural scientists endeavoured to quantify latent variables such as personality, intelligence, knowledge and attitudes [10, 20]. The science that grew out of this work is typically referred to as psychometrics [21], a term that appears to have been first used by Francis Galton in his 1879 paper *Psychometric experiments* [22], which he opens by stating that "Psychometry, it is hardly necessary to say, means the art of imposing measurement and number upon operations of the mind…" (p. 149). Other early key contributors include, e.g., Francis Edgeworth, Charles Spearman and Rensis Likert [20, 23]. The methodologies that resulted from these early scholars are commonly termed traditional, or classical test theory (CTT) [20].

The basic idea of CTT is that the observed score (*O*) can be decomposed into two main components, the true score (*T*) and an error score (*e*), formally expressed as $O = T + e$. The perhaps most obvious implementation of this expression is found in estimates of score reliability, which in general is expressed as the true score variance over the observed score variance, formally

$$\frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2}, \qquad (1)$$

where σ_T^2 is the true score variance and σ_e^2 is the error variance [21, 23]. Common applications of score reliability include coefficient alpha ("internal consistency") [24] and test-retest stability as assessed using the intraclass correlation coefficient of scores from two time points (under the assumption of no change in true score) [25]. Score reliability forms the basis for the CTT estimation of measurement uncertainty, which is expressed as the standard error of measurement (SEM),

$$SEM = SD\sqrt{1-r}, \qquad (2)$$

where SD is the standard deviation of the observed scores and *r* is score reliability [21].

Application of CTT in the development and testing of measurement instruments typically involves methods such as coefficient alpha, intraclass correlations, item-total correlation, inter-item correlations, associations with other variables, and principal component and factor analyses [21, 26]. These are all based on correlations, which is problematic if the purpose is to achieve measurement, since correlations concern associations rather than the extent to which numbers represent measures. This was noted as early as 1931 by Thurstone who, in the preface of his book *The reliability and validity of tests* remarked that:

"I do not believe that these correlational methods and particularly the reliability formulae have been responsible for much that can be called fundamental, important, or significant in psychology. On the contrary, the correlational methods have probably stifled scientific imagination as often as they have been of service. As tools in their proper place they are useful but as the central theme of mental measurement they are rather sterile" [27].

Additional caveats include that these correlational procedures are typically parametric, based on the Pearson product-moment correlation, which assumes quantitative normally distributed data. That is, they are not appropriate for ordered categorical item-level data [28, 29]. Furthermore, and related to its reliance on correlations, CTT is distribution dependent. This means that estimated psychometric properties cannot be generalized beyond the characteristics of the particular sample used in the analyses, which emphasizes the importance of representative samples and normal distributions of scores.

In CTT the raw summed total score is taken as a measure of the underlying latent variable, which implies linearity and a known unit of measurement. However, numerically coded response categories are ordered categorical data where the actual distances between categories remain unknown. Strictly speaking, such scores are therefore unsuitable for commonly employed traditional parametric statistical operations [29].

2.3. Rasch measurement theory

In the 1950s, Danish mathematician Georg Rasch proposed a novel approach to latent variable measurement [30], nowadays referred to as Rasch measurement theory (RMT). In

contrast to CTT, RMT is not based on correlations and does not take the observed total score to be a measure of the underlying variable. Instead, RMT is probabilistic rather than correlational in nature; as described by Rasch [30]:
”A person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one” (p 117).
In the dichotomous case, where item responses are scored either 0 or 1 (e.g., no/yes, disagree/agree, fail/pass), this means that the probability that a person will affirm or pass an item is a function of the location of the person (estimated “ability” measure) and the location of the item (estimated “difficulty” measure) on the same latent continuum. The more able the person is relative to the difficulty represented by the item, the more likely (s)he is to affirm or pass that item (and vice versa).

The basic Rasch model is expressed as the natural logarithm (*ln*) of the probability of success (a score of 1) over the probability of failure (a score of 0), which gives a log-odds unit (logit) representing the difference between the location (e.g., ability) of the person and the location (e.g., difficulty) of the item. Formally

$$ln\left(\frac{P_{ni1}}{1-P_{ni1}}\right) = \beta_n - \delta_i, \qquad (3)$$

where P_{ni1} is the probability of person *n* succeeding (scoring 1) on item *i*, β_n is the location of person *n*, and δ_i is the location of item *i*. Rasch’s basic model may also be expressed as

$$P_{ni1} = \frac{exp^{(\beta n-\delta_i)}}{1+exp^{(\beta n-\delta_i)}}, \qquad (4)$$

which probably is the more common form. As seen in Eq. (3) and (4) the probability of a certain response is governed by the distance between item and person locations. That is, if $\beta_n > \delta_i$, then P_{ni1} >0.5 and if $\beta_n < \delta_i$, then P_{ni1} <0.5, with 1 and 0 being the most likely outcomes, respectively. The point along the latent continuum where $\beta_n = \delta_i$, there is a 50/50 probability of either outcome; this is taken as the location (level of difficulty) of the item (δ_i). Since locations of both persons and items are expressed in logits, they have the same meaning across the continuum. However, locations are not absolute but relative, typically in relation to the average item location, which commonly is set at 0 logits.

In addition to the estimation of a separate location of each person and each item, the precision (standard error) of each location can also be estimated. For person locations the standard error is given by

$$\frac{1}{\sqrt{\sum_{i=1}^{L} P_{ni}(1-P_{ni})}}, \qquad (5)$$

for a total of *L* attempted items [31]. In contrast to CTT, this provides direct information on the measurement uncertainties associated with each individual person and item location. These standard errors are not constant but depend on the total score and vary along the range of measurement. In general, measurement uncertainty is greater for people with very low or high scores. However, its exact magnitude and pattern depends on the number and density of item or (in the polytomous case; see below) response category threshold locations, as well as on the alignment (targeting) of person and item/threshold locations.

The Rasch model for dichotomous response categories can be seen as a special case of the more general polytomous Rasch model (PRM). The PRM applies when possible responses or ratings consist of more than two ordered categories [32-36]. There are several versions of the PRM, of which the “rating scale model” (RSM) and the “partial credit model” (PCM) are the most common. While the RSM and PCM have been shown to be mathematically equivalent [34], the distinction lies in that the RSM assumes that response categories are the same and function the same way across all items, whereas the PCM does not [35]. The PRM takes the following general form:

$$P_{nix} = \frac{exp^{-\tau_{1i}-\tau_{2i}\cdots-\tau_{xi}+x(\beta_n-\delta_i)}}{\sum_{x'=0}^{m_i} exp^{-\tau_{1i}-\tau_{2i}\cdots-\tau_{x'i}+x'(\beta_n-\delta_i)}}, \qquad (6)$$

where P_{nix} is the probability of person *n* to score *x* on item *i*, τ_{xi} (x=1, 2, …m_i) are the thresholds that partition the latent continuum of item *i* into m_i+1 ordered categories, and *x* is the manifest item score. Thus, and analogous to the dichotomous case, τ_{xi} is the point at which there are equal probabilities of responding either in category *x* or category *x*–1 (given that the person scores in either of categories *x* or *x*–1). In the PRM, item locations are typically defined by the average of the threshold locations of each item.

Since ordered response categories are intended to reflect increasing (or decreasing) levels on the latent variable, examination of the locations of the estimated thresholds provides direct evidence as to whether response categories work as intended and represent an ordered hierarchy from less to more. That is, the increasing levels intended to be represented by each item's ordered response categories are expected to be reflected in the data and manifested as an expected ordering of thresholds [33, 37]. This is because each threshold is estimated as if it represented a dichotomous item (see above). The situation is therefore analogous to the estimation of dichotomous item locations that have an *a priori* expected hierarchy from less to more. Whether thresholds are ordered or disordered is thus a property of the data, and (as in the dichotomous case) does not necessarily impact tests of fit between the data and the model, which has led to some debate regarding the importance of threshold disordering [37, 38].

While RMT does not consider the observed total score a measure (but rather an ordered count), it is a sufficient statistic. This means that there is no other information than that embedded in the total score [39]. Specifically, a person's observed total score represents the sum of that individual's response probabilities for each item. Another unique feature of RMT is what Rasch referred to as specific objectivity [40], which means that person and item parameters can be separated. Since RMT is centered on the individual persons and items, Rasch set out the following requirement in the preface of his 1960 book [30]:
"It is further essential that comparisons between individuals become independent of which particular instruments – tests or items or other stimuli – within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class – "measuring the same thing" – independent of which particular individuals within a class considered were instrumental for the comparison" (p VII).
As further elaborated in his 1977 paper [40], by "objectivity", Rasch meant that any comparison of two people should be independent of what items are used (and vice versa for comparisons of items), and the qualifier "specific" refers to that the comparison is restricted to a specified frame of reference. Specifically, the frame of reference (*F*) was defined as: *F* = [*O*, *A*, *R*], where *O* is the object, *A* is the agent, and *R* is the range of possible reactions or outcomes [40]. The separation of person and item parameters is realized since the comparison of two person locations under the Rasch model is not dependent on any other parameters (e.g., item locations) or comparisons. For example, by simplifying and reducing to the interaction between single persons and items, it follows from Eq. (3) and (4) that the comparison of two persons (β_A and β_B, respectively) who have responded to the same item with location δ_i becomes (β_A – δ_i) – (β_B – δ_i), where δ_i cancels out and the comparison reduces to β_A – β_B. The situation is symmetrical for comparisons of item parameters.

RMT may be seen as a procedure that tests the extent to which items (and response categories) successfully have been able to map out the conceptual framework of the variable that they intend to represent. The estimated item (δ_i above) and threshold (τ_{xi} above) locations can be seen as representing markers on a latent "ruler". As such, and analogous to response category threshold ordering (see above), the hierarchical item ordering represents evidence as to the extent to which the empirical variable expression from less to more is in accord with conceptual substantive theory and clinical experience. That is, the empirical item hierarchy provides evidence as to how successful items are in operationalizing the latent variable as a theoretically or clinically coherent line of inquiry. Item hierarchies may also be useful from a clinical perspective. In rehabilitation settings, e.g., a well established hierarchy may serve as a means to help optimize and individualize interventions.

However, further quality control indices are also available to guide the measurement process through various tests of fit between empirical data and model expectations [31, 39, 41-43]. The accordance between data and the model can be assessed visually by inspection of empirical item responses relative to model expectations, represented by item characteristic curves (ICCs). A common quantification of this relationship in tests of model fit is the standardized fit residual (Z_{ni}), which has an expected value of 0 and represents the discrepancy between observed and expected item responses [31]. The basic fit residual expression can be formalized as:

$$Z_{ni} = \frac{X_{ni} - E[X_{ni}]}{\sqrt{V[X_{ni}]}}, \qquad (7)$$

where X_{ni} and $E[X_{ni}]$ are the observed and expected responses by person n to item i, respectively, and $V[X_{ni}]$ is the variance. Squaring and summing these residuals over all persons provides an overall approximate chi-square item statistic. In combination with other aspects of the data (e.g., descriptive statistics and reliability), these tests help diagnosing departures from model expectations.

RMT (as well as CTT) has two main requirements, unidimensionality and local response independence. Unidimensionality means that items should represent one common latent variable, and local response independence means that responses to one item should not be influenced by responses to one or several other items [39, 44]. These aspects are reflected in residual based tests of fit. Fit residuals have an expected value of 0 and a generally acceptable range between −2.5 to +2.5. In general, large negative (below −2.5 or so) and positive (above +2.5 or so) values signal local response dependence and multidimensionality, respectively. Graphically, negative fit residuals are observed as steeper empirical observations than the expected ICC, and vice versa for positive fit residuals. The matrix of correlations between standardized item fit residuals can also be examined regarding patterns and magnitudes, where relatively large correlations suggest local dependency [45].

Analyses of fit residuals can also provide a basis for assessing measurement invariance across subgroups of individuals, referred to as differential item functioning (DIF), which can be considered an additional aspect of fit. There is a range of approaches to examining DIF [46]. For example, analysis of variance (ANOVA) of fit residuals between subgroups of people along the measured variable [47]. Because the Rasch model can accommodate missing data, DIF can be further examined by adjusting for DIF by splitting the item into two new subgroup specific items [48]. For example, an item exhibiting DIF by gender would be split and analysed as two separate items, one for men and one for women. In the case of DIF detection in two or more items this procedure can be used to identify artificial DIF; if adjustment for

DIF in one item removes DIF in the second item, this suggests that the latter represented artificial DIF [47].

It is important to note that these and other tests of fit between data and the model primarily should be interpreted interactively and relatively rather than individually and in an absolute sense [31, 49]. For example, tests of fit are associated with multiple statistical hypothesis testing and corrections for this should be applied [50]. Furthermore, chi-square derived P-values should not be over interpreted because the distribution of standardized fit residuals is only approximately chi-squared. Instead, it is recommended to view chi-square values as relative order statistics with no strict cut-off values, where items associated with clearly larger chi-squared values than other items signal problems that need further consideration [31]. Similarly, it is recommended that residual correlations should be interpreted relative to the average matrix correlation [45]. Furthermore, various aspects of the analyses need to be accounted for and interpreted interactively, since anomalies may represent deviations from expectations that (in part) are due to less than obvious reasons. For example, multidimensionality may contribute to the occurrence of disordered thresholds as well as of DIF [51, 52].

3. The central role of rating scale based outcome measures in clinical research

Rating scale based instruments constitute the main type of clinical trial outcome measures in a wide range of clinical specialties, including neurological disorders such as PD. This means that scores from such instruments often represent the central dependent variable upon which inferences and decisions are made regarding the usefulness of therapies. Therefore, such instruments often play a central role in high stakes decision making that ultimately impacts the health and lives of ill people, since the quality of clinical study results and inferences is directly dependent on the quality of the instruments used to collect data [53].

The central role of outcome measures was also emphasized in a recent UK study conducted among people with PD, their carers, families and friends, and health care professionals to identify research priorities for PD management [54]. Along with finding a cure and improved symptom control, the development of better methods to monitor treatment responses was ranked among the top 10 research priorities. Another study [55] investigated how people with PD ranked the importance of experiencing therapeutic improvements in 10 predefined areas. Results suggested that ADL together with walking and slowness of movements were considered the most important areas from the patients' perspective.

The aspects discussed in the two previous paragraphs illustrate several important points. First, if we as clinicians and clinical investigators take our patients and our studies seriously, we also need to be serious about our outcome measures. Unless they are treated with full scientific rigour, advances in the clinical sciences will be hampered and opportunities to improve patient care may be lost. Second, outcome measures do not only need to satisfy more or less arbitrary standard heuristics regarding their psychometric qualities, but they also need to be inherently meaningful in a manner that links qualitative and quantitative aspects with reference to a common linear continuum. In other words, in order to be meaningful and meet their purpose as clinical outcome measures, they need to represent (a) measures of the variables they intend to quantify; (b) the variables that are important and meaningful to patients, their families and health care professionals; and (c) the variables that therapies are intended to target. These ideas are largely in line with those of the United States' Food and Drug Administration, as first articulated in their guidance for industry regarding use of patient-reported outcome measures to support labelling claims [56, 57] and more recently in their Clinical Outcome Assessment Qualification Program [58].

3.1. An illustrative example

To illustrate the central role of rating scale based outcome measures in clinical investigations, key aspects of a typical double blind, placebo-controlled randomized controlled trial (RCT) will be reviewed. [1] The RCT was a post-marketing multicentre phase IV clinical trial that ran for 13 weeks and concerned an adjunct drug to be used together with levodopa in people with PD who experience wearing-off, i.e., episodes with insufficient efficacy of oral levodopa and increased PD symptomatology before the onset of the next scheduled levodopa dose. The primary objective was to determine the effect of the combination therapy on ADL. The primary outcome measure was the UPDRS II (ADL).

The UPDRS II was introduced in 1987 [15] and has since become the most widely used means of assessing ADL outcomes in clinical PD research, as well as in clinical practice [59, 60]. In a recent review commissioned and conducted by the MDS [60], it was deemed "recommended" for measuring disability in PD. The UPDRS II consists of 13 items that are intended to be completed as part of a standardized clinical interview where the examiner (typically a neurologist or specialized nurse) records historical patient-reported information by selecting one of five response categories available for each item (Table 1). Each item is scored from 0 ("normal" or absence of problems) to 4 ("severe" or inability to perform task), and item scores are summed to provide a total score that is taken to represent the person's level of ADL limitation, from 0 (no limitations) to 52 (severe limitations).

[1] Reference to the study or the name of the tested drug is not provided since this is irrelevant for the purpose of this discussion.

Table 1
Unified Parkinson's Disease Rating Scale, part II (activities of daily living).

Items: Number	Contents
1	Speech
2	Salivation
3	Swallowing
4	Handwriting
5	Cutting food and handling utensils
6	Dressing
7	Hygiene
8	Turning in bed and adjusting bed clothes
9	Falling (unrelated to freezing)
10	Freezing when walking
11	Walking
12	Tremor
13	Sensory complaints related to parkinsonism

Response categories: [a] Score	Description
0 [b]	Normal
1	Slight but definite excess of saliva in mouth; may have nighttime drooling
2	Moderately excessive saliva; may have minimal drooling
3	Marked excess of saliva with some drooling
4	Marked drooling, requires constant tissue or handkerchief
0 [c]	Normal
1	Somewhat slow, but no help needed
2	Occasional assistance with buttoning, getting arms in sleeves
3	Considerable help required, but can do some things alone
4	Helpless

[a] Response categories are unique for each item but intend to represent an ordered scale ranging from "normal" (or absence of problems) to "severe" (or inability to perform task), that is scored by successive integrals from 0 to 4 so that higher scores represent more activity limitations. Item scores are summed to provide a total score that is taken as a measure of the person's level of activity limitation.
[b] Example from item 2 (Salivation).
[c] Example from item 6 (Dressing).

Figure 2 illustrates the basic study design and results. Participants were randomly allocated either to receive the active adjunct drug or to the control group receiving inactive placebo, in addition to levodopa. Clinical assessments were conducted at three time points: at baseline (yielding UPDRS II scores of 23.5 and 24.7 in the active and placebo groups, respectively), after 5 weeks (not illustrated in Figure 2) and at the end of the trial (13 weeks after baseline). At the end of the study, the active group had improved their UPDRS II scores by a mean of 2.3 whereas the improvement was 0.7 in the placebo group; a difference of 1.6 UPDRS II points. This difference between the two study arms was found statistically significant at P<0.0001, and it was therefore concluded that the adjunct therapy had good efficacy in terms of ADL.

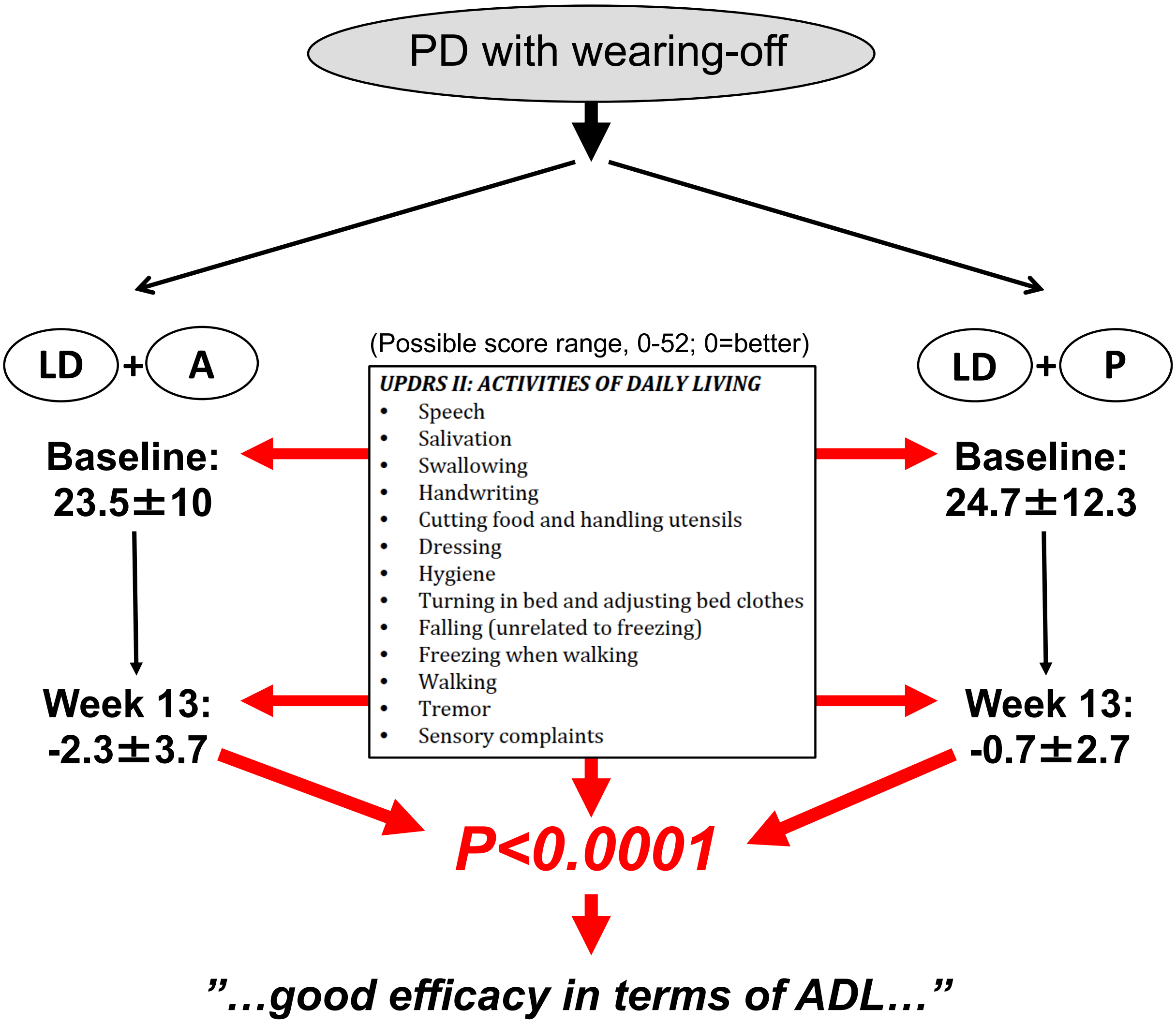


Fig. 2. Schematic illustration of the design and results of a randomized controlled phase IV trial of an adjunct active drug (A) or placebo (P) used together with levodopa (LD) in people with fluctuating Parkinson's disease (PD) to determine the effect on activities of daily living (ADL) using the UPDRS II as the primary outcome measure.

Several related and central questions arise from the review of this RCT:

- Can data and results be interpreted in a meaningful way?
- Are UPDRS II total scores valid representations of their target variable?
- How does the UPDRS II perform as an ADL measurement instrument in PD?

These questions are not unique for the RCT reviewed here, but would apply to any study that relies on rating scale derived data to represent central endpoint variables.

4. Material and methods

The questions arising from the study reviewed in section 3.1 were addressed by (i) a review of relevant literature, (ii) consideration of the conceptual basis and properties of UPDRS II items as operationalisations of the latent activity performance variable, and (iii) by examination of its psychometric properties using empirical data from people with PD, but otherwise unrelated to participants in the RCT reviewed above. Data were analysed using both CTT and RMT methodologies. CTT analyses are provided for two main reasons. First, this is the approach by which the UPDRS II has been assessed in the past and upon which it has gained status as a "recommended" scale for measuring ADL outcomes in PD [60]. Second, it will provide a point of reference for the results gained from the subsequent RMT analyses, which will be useful since the UPDRS II does not appear to have been examined by use of RMT before.

The empirical data used to examine the psychometric and measurement properties of the UPDRS II were taken from several studies conducted over the past decade. All assessments were conducted by experienced neurologists or nurses specialized in PD and trained in the use of the UPDRS. The sample consisted of 675 people with PD (Table 2) representing all five stages of PD according to Hoehn & Yahr [61]. There were no missing item responses. The original studies were conducted in accordance with the Declaration of Helsinki and all participants provided informed consent.

Table 2
Sample characteristics (n=675).

Age (years), mean (SD; min – max)	63.8 (10; 32 – 88)
Men / women, n (%)	410 (61) / 265 (39)
Time since diagnosis (years), mean (SD; min – max)	9.5 (7.2; 0.3 – 44)
Hoehn & Yahr stage of PD, median (q1 – q3; min – max) [a]	III (II – IV; I – V)

[a] Stage I, unilateral disease; Stage II, bilateral disease without postural instability; Stage III, bilateral disease with postural instability, somewhat restricted in activities but capable of leading an independent life, mild to moderate disability; Stage IV, severely disabled, still able to walk and stand unassisted but markedly incapacitated; Stage V, confinement to chair or bed unless aided [61].
SD, standard deviation; PD, Parkinson's disease; UPDRS, Unified PD Rating Scale.

CTT analyses were conducted in two ways: (i) the traditional or "default" approach [26, 62] that is used in a majority of CTT based psychometric work, including previous studies on the UPDRS II [60]. In this approach, correlations (including computation of coefficient alpha) are parametric, and dimensionality is assessed using principal component analysis (PCA) of a Pearson-based correlation matrix; (ii) a methodologically more appropriate approach [28, 63-65] based on polychoric/polyserial correlations and minimum rank factor analysis (MRFA). In both instances, three criteria were considered fӧr determination of the number of dimensions in data based on PCA and MRFA: Kaiser's eigenvalue >1 criterion, screeplots, and parallel analysis [63, 65, 66]. Other criteria include corrected item-total correlations ≥0.3-0.4, reliability (alpha) ≥0.8, and normal distribution of scores [21, 49]. CTT analyses were conducted using R (version 3.4.0) with the "psych" package (version 1.7.5) (www.r-project.org) and FACTOR (version 10.8.02) [63, 67].

RMT analyses were conducted as described in section 2.3 and focused on the following aspects: targeting, reliability, model fit, rating scale response category functioning, and DIF by gender, age, and time since diagnosis. Subgroups for analyses of DIF by age and time since diagnosis were defined according to their respective median values. RMT analyses were conducted according to the PCM version of the PRM using the RUMM2030 program (RUMM Laboratory Pty Ltd., Perth, WA, Australia), with the sample divided into ten class intervals (subgroups with similar levels according to their estimated locations). Bonferroni adjustments for multiple null hypothesis testing were applied (alpha level of significance, 0.05) [50, 68].

5. Results and discussion

5.1. Can data and results be interpreted in a meaningful way?

To begin addressing this issue one may first consider what the main result, a mean difference of 1.6 on the UPDRS II total score (i.e., about 3% of the possible 0-52 score range) actually means, and whether it represents a clinically meaningful or interesting difference. This question cannot be answered readily from within the data reported in the RCT. However, some investigators have attempted to establish the minimally important difference (MID) for the UPDRS II. That is, to determine the minimal degree of change in UPDRS II scores that can be considered to be of clinical significance. This was done by relating changes in UPDRS II scores to clinician's global impression of improvement. One such study suggested that improvements of about 2 UPDRS II raw total score points may represent the MID among people in stages I-II according to Hoehn & Yahr, and for people in Hoehn & Yahr stages II-III the MID was estimated to about 3 points [69]. Similar or larger values have been reported by other authors [70-72]. These estimates are associated with uncertainties, and 95% confidence interval (CI) widths have ranged from about 1 to 4.3 and in some cases they included a change score of zero [69-71]. However, it remains unknown what these differences actually mean in practical terms, i.e. how people with a certain magnitude of improvement (or deterioration) actually have benefitted. Returning to the outcomes from our example RCT in section 3.1, it is thus unknown whether the observed average difference of 1.6 between treatment arms is a clinically meaningful one, although most available evidence reviewed above would challenge this.

5.2. Are UPDRS II total scores valid representations of their target variable?

A central question regarding score interpretation and meaning is what variable scores represent. The extent to which scores can be interpreted as representing variations in its target variable depends in part on the definition according to which the variable was operationalized through its items when the scale was developed. In the case of the UPDRS II, the developers provide no conceptual reasoning or variable definition underpinning item selection [15]. Therefore, in order to address this issue in relation to an established conceptual framework, UPDRS II items were linked to the World Health Organization's International Classification of Functioning, disability and health (ICF) [16]. The ICF provides a conceptualisation and classification of different components of health including biological, individual, and social perspectives. Its main part defines functioning and disability, which consist of two components, (i) Body structures (i.e., anatomical integrity) and Body functions (i.e., symptoms and signs, typically refered to as impairments), and (ii) Activities (i.e., the execution of tasks or actions) and Participation (i.e., involvement in life situations). In order to be valid representations of their target variable, UPDRS II items are therefore expected to be linked to Activities according to the ICF.

Conceptual review and linking of items to ICF components revealed that six of the 13 UPDRS II items (items 4-8 and 11) represent Activities, whereas seven items (1-3, 9, 10, 12, and 13) represent Body functions (impairments). Similar results have been reported by Hariz and collaborators [73] who also conducted a conceptual review but without linking to the ICF. In the absence of a clear variable definition [15], this suggests that the UPDRS II does not represent a common unidimensional variable. This has at least two implications. First, it is unknown what a certain score means and, second, score comparisons (between or within people) lack clinical meaning since the degree of representation of the respective dimensions in a particular total score is unknown. That is, the same total UPDRS II score does not necessarily represent the same degree of activity limitation and, conversely, two people with the same degree of activity limitation may well have different UPDRS II scores due to

symptomatic differences. For example, consider two patients whose ADL disabilities are the same, as reflected by a score of, say, 10 on items representing activities, but one of them has relatively pronounced impairments manifested as a score of 12 on the impairment items whereas the other has an impairment score of 5. Their total UPDRS II scores are 22 and 15, respectively. Conversely, two persons with the same total UPDRS II score may have quite different activity and impairment scores. Both situations inhibit proper interpretation and meaning of the total score and may also have different therapeutic implications. Curiously, the authors of the MDS review of disability scales appear to have been aware of this when they concluded that the UPDRS II is "recommended" as an ADL outcome measure in PD [60]:
"The major weakness is that the 13 items of the UPDRS-ADL do not all assess disability…" (p 1460).

Although the UPDRS II was the primary ADL outcome measure in our example RCT above, the trial also included two other scales purporting to represent ADL, the Schwab & England ADL scale [15] and the ADL scale of the 39-item PD Questionnaire (PDQ-39) [74]. However, in contrast to the UPDRS II results according to these secondary outcome measures did not show any differences between the two treatment arms, thus adding further concerns to the issue of validity and clinical interpretability of scores.[2]

5.3. How does the UPDRS II perform as an ADL measurement instrument in PD?

Measurement is strongly linked to substantive theory and conceptualisation [75, 76]. Therefore, consideration of the measurement properties of the UPDRS II will, per definition, also relate to the two initial questions addressed above. First, results from CTT based analyses are presented, followed by those according to RMT.

5.3.1. Classical test theory

Results from the analyses according to the CTT tradition are summarized in Table 3 and Figure 3.

[2] This is not to say that either of the two alternative ADL scales are preferable over the UPDRS II. That would require a thorough review and analysis of these scales as well, which is beyond the score of this paper.

Table 3
Descriptive and psychometric statistics of the UPDRS part II (activities of daily living) according to classical test theory (n=675).

Descriptive statistics	
Total score median (q1 – q3)	7 (0 – 16)
Total score min – max	0 – 45
Total score mean (SD)	9.7 (10.2)
Total score skewness (SE)	1.0 (0.09)
Total score kurtosis (SE)	0.26 (0.19)
Total score floor-/ceiling effects, % [a]	30.7 / 0
Item median scores, min – max	0 – 1
Item mean scores, min – max	0.36 – 1.18
Item SD, min – max	0.72 – 1.29
Item skewness, min – max	0.73 – 2.24
Item kurtosis, min – max	-0.76 – 4.47
Psychometric CTT statistics	
Corrected Pearson based item-total correlations, min – max [b]	0.52 – 0.88
Corrected polyserial based item-total correlations, min – max [b]	0.63 – 0.92
Cronbach's alpha [c]	0.94
Cronbach's alpha when item deleted, min – max [d]	0.93 – 0.94
Ordinal alpha [e]	0.96
Ordinal alpha when item deleted, min – max [d]	0.96 – 0.96
SEM (based on Cronbach's alpha) [f]	2.50
SEM (based on ordinal alpha) [f]	2.04
Pearson based PCA	
C1 loadings, min – max	0.57 – 0.91
C1 / C2 eigenvalues	7.80 / 1.03
C1 / C2 eigenvalues from parallel analysis [g]	1.28 / 1.22
Polychoric based MRFA	
F1 loadings, min – max	0.65 – 0.95
F1 / F2 % common variance explained	78.80 / 6.30
F1 / F2 % common variance explained from parallel analysis [g]	27.60 / 23.60

[a] Percentage of persons with the lowest (floor) and highest (ceiling) possible total score; suggested to be acceptable if ≤15-20% [49].
[b] The correlation between item score and the total score excluding that item; suggested to be acceptable if ≥0.30–0.40 [49].
[c] Traditional parametric estimate of score reliability [24]; suggested to be acceptable if ≥0.80 [21].
[d] Should not increase compared with alpha for the total score [21].
[e] An estimate of score reliability based on polychoric correlations [64]; suggested to be acceptable if ≥0.80 [21].
[f] SEM = SD x √(1 – reliability) [21].
[g] Parallel analysis based on the 95th percentile of eigenvalues from 500 random correlation matrices obtained by empirical data permutation [65].
UPDRS, Unified Parkinson's Disease Rating Scale; q1-q3, first-third quartile; SD, standard deviation; SE, standard error; SEM, standard error of measurement; PCA, principal component analysis; C, component; MRFA, minimum rank factor analysis; F, factor.

Overall, the UPDRS II appears to behave well and meet most CTT criteria. For example, scores appear reliable and associated with reasonable measurement uncertainties (with SEM values representing about 4-5% of the possible total score range). Furthermore, items appear to represent a single common latent variable as suggested from corrected item-total correlations as well as from PCA and MRFA analyses according to all three applied criteria. It is evident, however, that the scale appears to represent greater activity limitations than those experienced by the sample since almost one third of the people had a total score of zero (i.e., a floor effect) and mean and median scores represent only about 13 and 19 per cent of the total score range, with a similar pattern among item level scores (Table 3). This indicates that the CTT assumption of normally distributed data is not met. However, in line with the conclusions by the MDS [60], these results suggest that the UPDRS II generally would be considered a psychometrically sound tool for measuring ADL outcomes in people with PD.

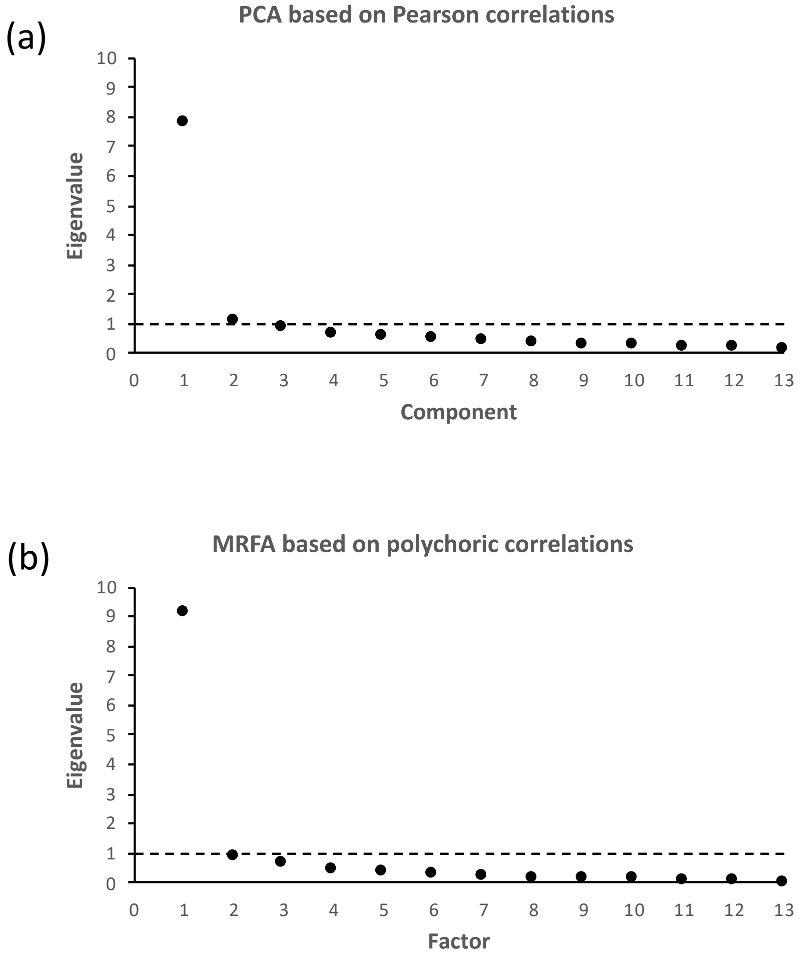


Fig. 3. Scree plots of the eigenvalues (y axes) for components (x-axis, panel a) and factors (x-axis, panel b) identified by principal component analysis (PCA; panel a) and minimum rank factor analysis (MRFA; panel b) of item-level UPDRS II data. The dashed horizontal line indicates the cut-off point for determination of the number of components and factors according to the eigenvalue >1 criterion.

Although not uncontroversial, the SEM has been suggested as a distribution based estimate of the MID [77, 78]. Based on this, the SEM estimated from the data presented here would suggest an MID of 2-3 for the UPDRS II total score (Table 3). This is similar to previously reported MID estimates for the UPDRS II (see section 5.1) but above the 1.6 average group difference in the study reviewed in section 3.1. Again, this would challenge the clinical meaningfulness of the RCT results.

5.3.2. Rasch measurement theory

Table 4 and Figure 4 illustrate the targeting and measurement precision of the UPDRS II, as derived based on the Rasch model.

Table 4
Targeting and reliability of the UPDRS part II (activities of daily living) according to Rasch measurement theory (n=675).

	Total score	Activity score	Impairment score
Targeting [a]			
Person location, mean (SD)	-2.37 (1.88)	-2.69 (2.53)	-1.82 (1.13)
Person location, min – max	-4.79 – 1.97	-5.42 – 6.28	-3.17 – 0.82
Item location, mean (SD)	0 (0.5)	0 (0.73)	0 (0.36)
Item threshold location, min – max	-2.75 – 2.99	-3.68 – 5.39	-1.65 – 1.84
Reliability			
Person separation index [b]	0.82	0.86	0.51
Number of distinct strata of people [c]	3.2	3.7	1.7

[a] Relative to the mean item logit location (i.e., zero).

[b] Conceptually analogous to Cronbach's alpha [79].

[c] Number of statistically distinct groups of people (separated by ≥3 standard errors) that can be distinguished [35].

UPDRS, Unified Parkinson's disease Rating Scale; SD, standard deviation.

It is seen that the scale represents a continuum from lower to higher levels (ranging 5.75 logits, from about -2.75 to 3 logits; Fig. 4a, lower panel). This range is similar to that found in the sample (ranging approximately 6.76 logits, from about -4.79 to 1.97 logits; Fig. 4a, upper panel). However, the scale represents more severe states than those observed in the sample. For example, the sample was located an average of about 2.4 logits below the average item location (set at 0). As a consequence, people located at the lower range are measured with relatively poor precision. This is illustrated by the superimposed 95% CI of person measures at various locations along the measurement continuum (Fig. 4a).

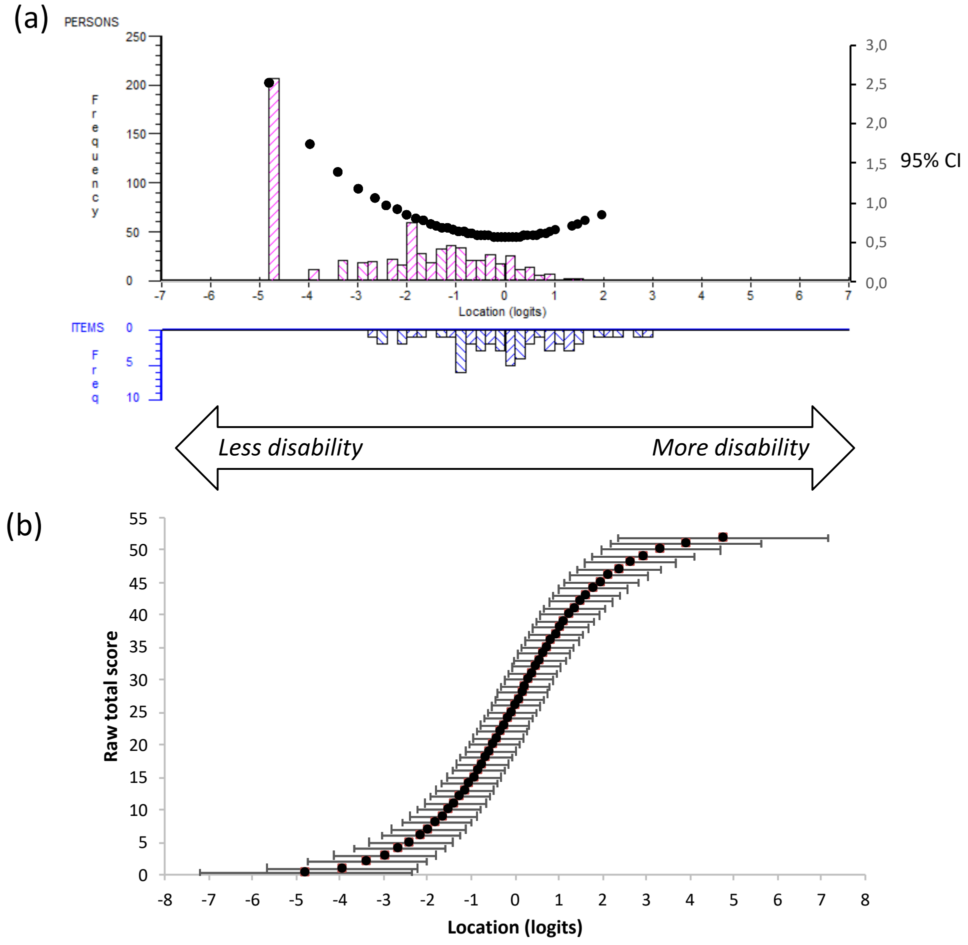


Fig. 4. (a) Distribution of locations of persons (upper panel) and UPDRS II response category thresholds (lower panel) on the common logit metric (*x*-axis; negative values = better). Superimposed black dot represent the uncertainties (95% CIs, right y-axis) of the observed person locations along the measurement range. (b) Relationship between raw total UPDRS II scores (y-axis) and their implied linear locations on the logit metric (x-axis) with estimated lower and upper 95% CI limits of uncertainty (±1.96*SE*; represented by horizontal error bars) across the full range of all possible UPDRS II raw total scores.

For example, the best point of measurement is seen around 0 logits (a 95% CI of ±0.55 logits), whereas measurement uncertainty is considerably larger (a 95% CI of ±0.96 logits) at the average person location (2.4 logits). This is further illustrated in Figure 4b, which depicts the estimated relationship between each possible raw total score and its implied linear logit measure, including the respective measurement uncertainties. Taking this information into account, it is for example seen that the average person location (2.4 logits) represents a raw total score of 5 with an uncertainty ranging between total scores of 2 and 11. Refering back to the results from the RCT reviewed in section 3.1, raw total UPDRS II scores at baseline were about 24 in both treatment arms, which corresponds to a linear measure of -0.15 logits. The uncertainty at this location has lower and upper 95% CI limits of -0.71 and 0.41 logits, which correspond to raw total scores of 17 and 31, respectively. Clearly, this yields further doubt about the clinical significance of a raw total score group difference of 1.6.

Tests of fit between data and the Rasch model are reported in Table 5 and Figure 5. In general, tests of fit can be seen as an investigation into the extent to which items successfully represent the intended target concept. Table 5 suggests that eight of the 13 UPDRS II items do not exhibit statistical fit to the model according to chi-square derived P-values.

Table 5

Item locations and fit statistics of the UPDRS part II (activities of daily living) according to Rasch measurement theory (n=675).

	Total score					Activity score					Impairment score				
Item [a]	Location (95% CI) [b]	Fit Residual [c]	Chi-2	P-value [d]	DIF [e]	Location (95% CI) [b]	Fit Residual [c]	Chi-2	P-value [d]	DIF [e]	Location (±95% CI) [b]	Fit Residual [c]	Chi-2	P-value [d]	DIF [e]
1bf	0.06 (0.12)	1.32	17.65	0.040	D	na	na	na	na	na	-0.18 (0.11)	-1.85	24.02	**0.004**	D
2bf	0,38 (0.13)	2.39	32.97	**<0.001**	A	na	na	na	na	na	0.16 (0.12)	-0.18	10.00	0.351	A
3bf	1.08 (0.14)	**2.65**	27.95	**0.001**	-	na	na	na	na	na	0.70 (0.13)	-1.31	21.04	0.012	-
4a	-0.89 (0.10)	-1.63	14.10	0.119	-	-1.00 (0.13)	**3.94**	17.17	0.046	-	na	na	na	na	na
5a	-0.16 (0.12)	**-5.20**	33.27	**<0.001**	-	0.19 (0.14)	**-2.57**	16.74	0.053	-	na	na	na	na	na
6a	-0.30 (0.12)	**-5.76**	51.69	**<0.001**	-	0.05 (0.15)	**-3.25**	20.22	0.017	-	na	na	na	na	na
7a	0.47 (0.13)	**-5.23**	44.27	**<0.001**	D	1.22 (0.16)	**-2.64**	14.25	0.114	-	na	na	na	na	na
8a	-0.34 (0.11)	-1.45	16.26	0.062	-	-0.15 (0.14)	2.48	10.39	0.320	-	na	na	na	na	na
9bf	0.14 (0.11)	1.93	13.53	0.140	G	na	na	na	na	na	-0.14 (0.10)	-1.62	15.02	0.090	G, D
10bf	0.01 (0.11)	-1.92	19.57	0.021	D	na	na	na	na	na	-0.23 (0.09)	**-3.57**	28.34	**0.001**	G, D
11a	-0.49 (0.12)	**-3.70**	28.71	**0.001**	-	-0.31 (0.14)	1.42	8.13	0.522	-	na	na	na	na	na
12bf	-0.25 (0.11)	**6.60**	107.04	**<0.001**	D	na	na	na	na	na	-0.40 (0.10)	**3.18**	31.11	**<0.001**	D
13bf	0.31 (0.11)	**5.30**	73.19	**<0.001**	-	na	na	na	na	na	0.11 (0.10)	0.85	17.93	0.036	-

[a] “a” and “bf” indicates whether items conceptually map to the Activity or Body function (impairments) domains of the ICF. See Table 1 for item descriptions.

[b] Item locations are the mean of each item’s response category threshold values expressed in linear log-odds units (logits); 95% CIs are 1.96*SE*.

[c] Standardized fit residuals summarise the deviation of observed from expected responses, and are recommended to range about -2.5 to +2.5 [31].

[d] Bonferroni corrected statistically significant deviations, suggesting item misfit, are bold.

[e] A, DIF by age; D, DIF by disease duration; G, DIF by gender after accounting for artificial DIF.

UPDRS, Unified Parkinson’s Disease Rating Scale; CI, confidence interval; DIF, differential item functioning; na, not applicable.

Inspection of the chi-square statistics from a relative point of view (Fig. 5a) suggests that items 12 and 13 (both representing impairments rather than activities) represent the main deviation from model expectation. The graphical representation of fit between data and model expectations for these two items supports this (Fig. 5b and 5c) and the pattern of deviation suggest that these items may represent a different variable than that captured by the scale as a whole. Interestingly, this observation appears to generalize since standardized fit residuals suggest a pattern in that items representing impairments primarily appear to signal multidimensionaility, whereas activity items tend to show signs of local response dependence (Table 5).

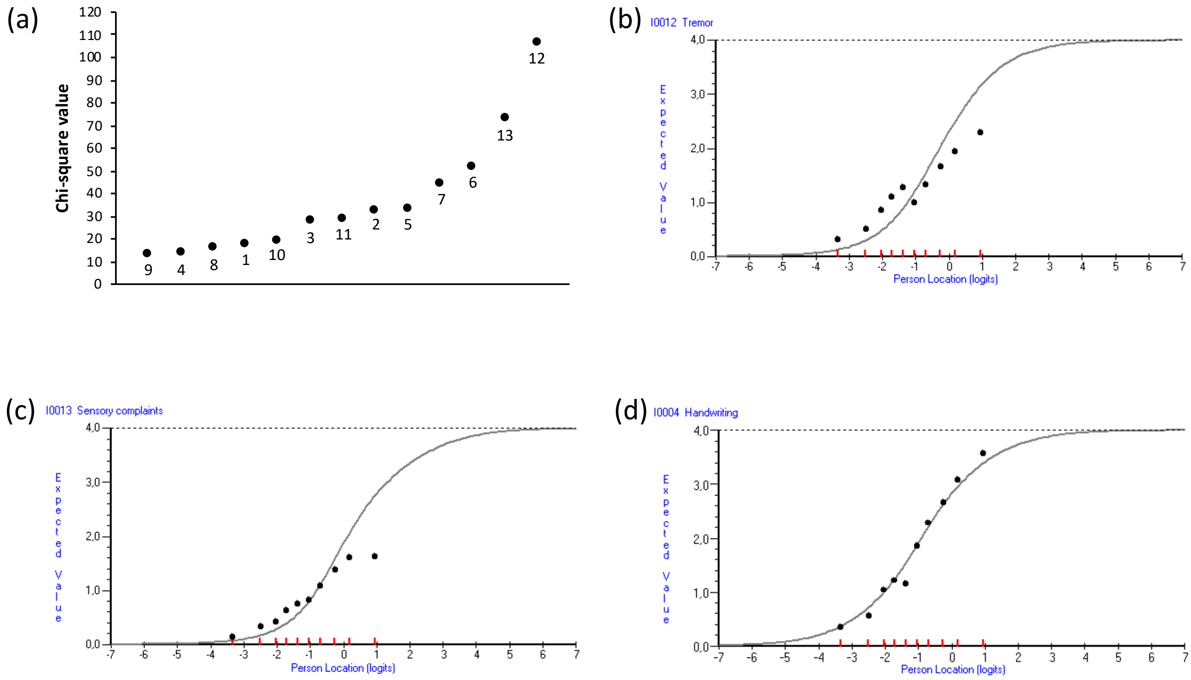


Fig. 5. Graphical representation of tests of fit between UPDRS II items and the Rasch model. Panel (a) Item chi-square statistics plotted in ascending order with numbers representing item numbers. Panels (b) through (d) display item characteristic curves (ICCs) of expected (grey curves) and observed (black dots) item responses (y-axis) for the 10 class intervals (subgroups of people with different locations; n=41-59 per class interval) along the outcome continuum (x-axis). Panels (b) and (c) display the worst fitting items (items 12 and 13, respectively). For comparison, panel (d) represents an item with relatively good fit (item 4).

Inspection of the matrix of standardized fit residual correlations supported this (data not shown). Activity items tended to display larger correlations (up to 0.47) among one another than impairment items did (up to 0.27), and activity items loaded positively whereas impairment items loaded negatively on the first principal component following a PCA of the standardized fit residuals (data not shown). A series of t-tests comparing the person locations estimated from activity and impairment items, respectively, showed that person locations differed significantly for about 10% of the sample.

Six of the 13 items were also found to exhibit DIF (Table 5), suggesting that they function differently across age, gender or disease duration groups. It is also noted that all but one (item 7) of the six items with DIF represent impairments rather than activities.

Assessment of the empirical functioning of response categories showed that these did not work as intended with four items (Fig. 6a). Specifically, two patterns of disordered thresholds emerged, which seem to suggest that either three (items 2, 3 and 10; Fig 6b) or two (item 9; Fig 6c) categories would be preferable to the current five response categories. The observed disordering may be due to difficulties in distinguishing between five levels of salivation, swallowing, falling and freezing, respectively, or overlapping or undistinct response category wording. For instance, it is unclear how to clearly separate between “minimal” and “some” drooling (item 2; Table 2, Fig. 6b). Indeed, these observations appear to make clinical sense. For instance, it is unclear how to clearly separate between “minimal” and “some” drooling (item 2; Table 2, Fig. 6b). Rewording of response categories may thus be one way of improving these items. However, it is also noted that one reason for disordered thresholds may be that the item represents multidimensionality. From this perspective, it is noted that all four cases of threshold disordering represented impairment items.

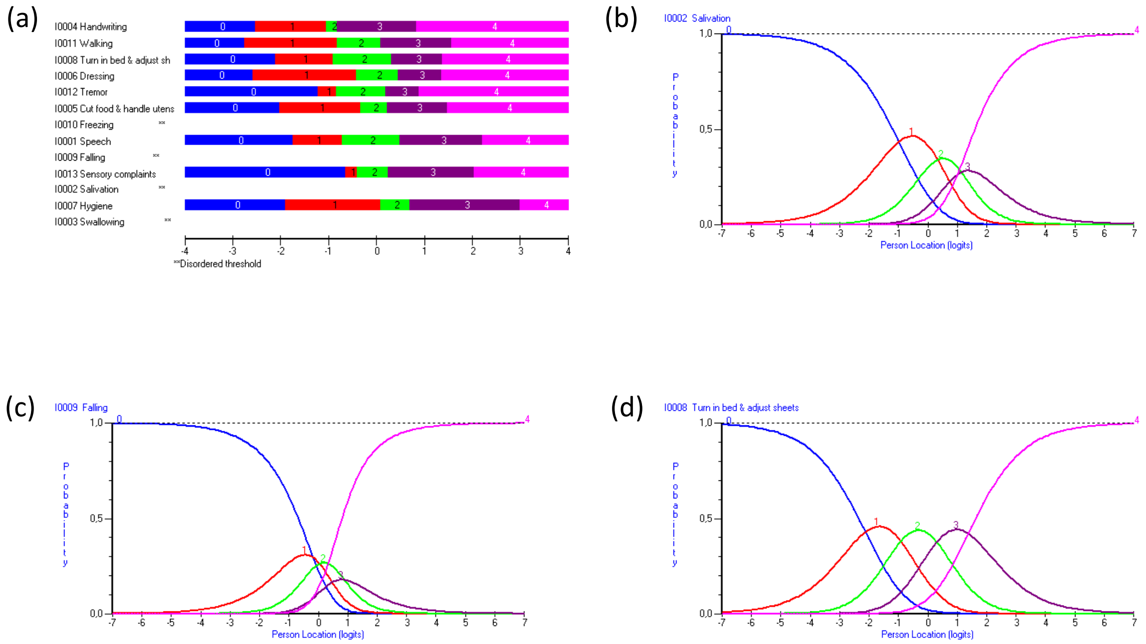


Fig. 6. Response category functioning for UPDRS II items. Panel (a) provides an overview of all items ordered hierarchically from the top. Intersections between color bars represent response category thresholds. Instances of disordered thresholds are blank since disordering prevents this type of graphical representaation. Panels (b) through (d) provide a more detailed picture, where each colored category probability curve represenst the probability (y-axis) of responding in that response category relative to various estimated person logit locations (x-axis). Panels (b) and (c) illustrate patterns of disordered thresholds for items 2 and 9, respectively. For comparison, panel (d) illustrates an item without disordered thresholds (item 8).

In Figure 6a, items have been ordered by their locations (the average of each item’s threshold values) to illustrate their hierarchical structure from the “easiest” or most commonly endorsed (item 4, handwriting) to the “hardest” (item 3, swallowing). Figure 7a also illustrates the item hierarchy but in more detail, including the uncertainties (95% CIs) of each item location.

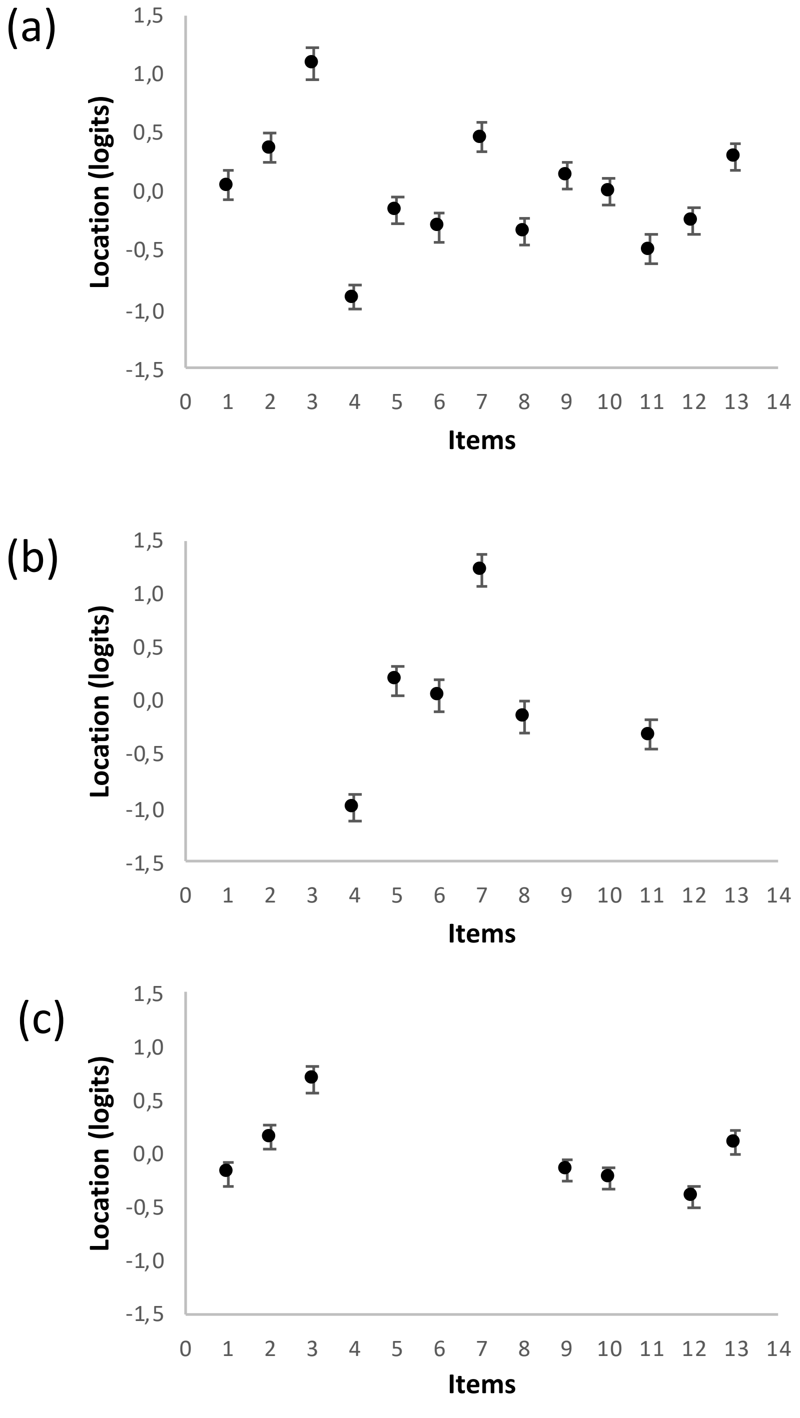


Fig. 7. Hierarchical item ordering with item locations on the y-axis (negative values = better) and item numbers on the x-axis (see Table 1 for item descriptions). Error bars are uncertainties (±95% CIs). Panel (a) represents the total UPDRS II score, whereas panels (b) and (c) displays the item hierarchies when analyzing activity (b) and impairment (c) items separately.

Taken together, in contrast to CTT, conceptual and RMT based evidence suggest that the UPDRS II does not provide a useful means for measuring ADL outcomes in PD. Instead, it appears to represent two different constructs, i.e., Activities (items 4-8 and 11) and Body Functions/impairments (items 1-3, 9, 10, 12, and 13). The main implication of this is that any interpretation of results from the UPDRS II (whether raw total scores or linear logit measures) is ambiguous.

5.4. Can the UPDRS II be improved?

To understand the extent to which the two embedded activity and impairment item sets may provide useful measures of their respective constructs, data were reanalysed using RMT methodology but treating the two item sets as two separate scales.

Table 5 and Figure 8 illustrate targeting and measurement precision of the two new UPDRS II derived activity (Fig. 8a) and impairment (Fig. 8b) scales. It is seen that activity items represent a wider range of measurement (9.07 logits, from about -3.68 to 5.39 logits; Fig. 8a), whereas the impairment items are narrower (3.5 logits, from about -1.65 to 1.84 logits; Fig. 8b) compared to the original UPDRS II (Fig. 4). These effects are also mirrored in the estimated person measures and reliabilities (Table 5) where, for example, the 6-item activity scale actually exhibits somewhat better reliability than the original 13-item UPDRS II and still is able to separate between 3 and 4 statistically distinct strata of people. In contrast, the 7-item impairment scale exhibits a substantially reduced reliability.

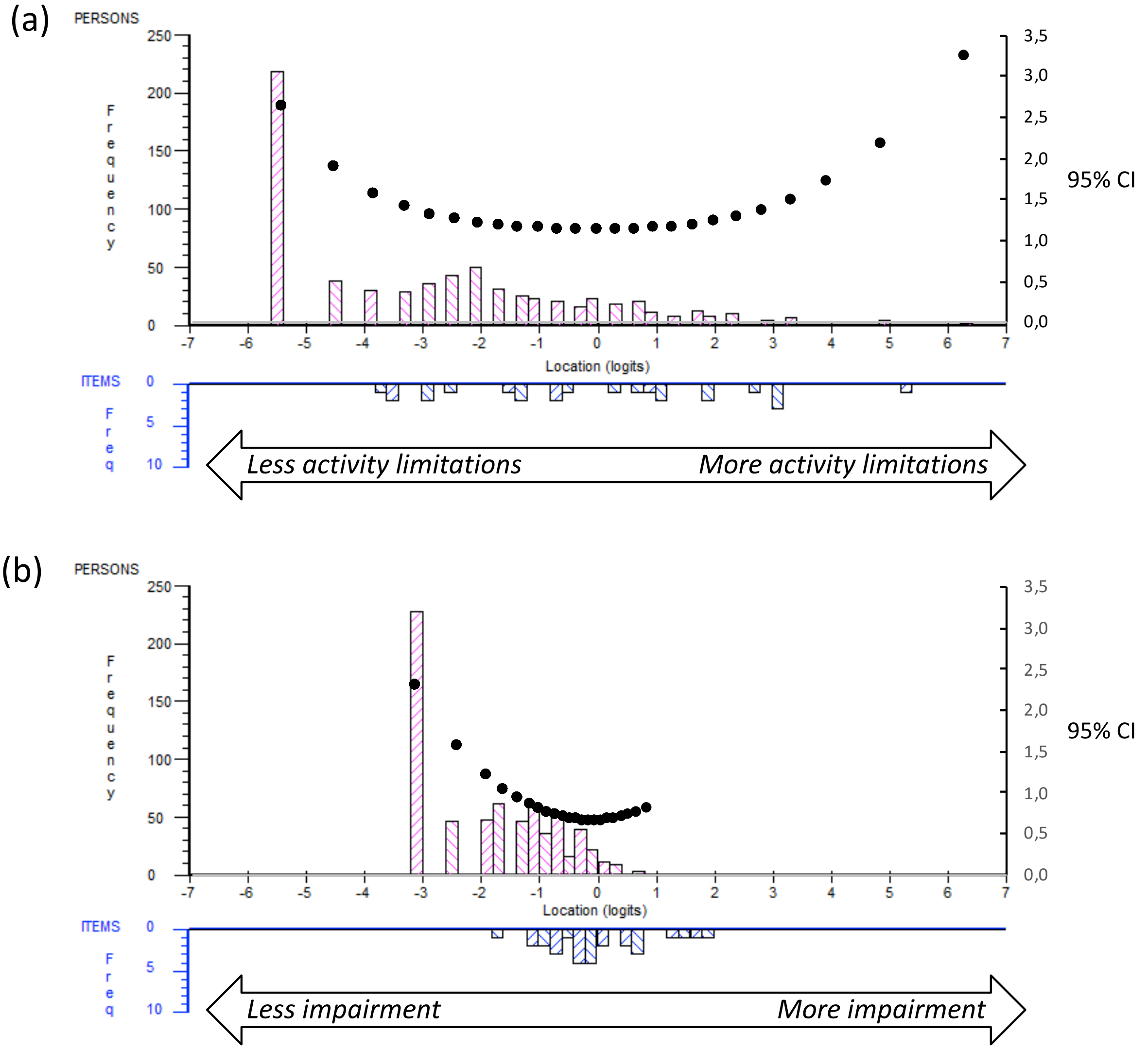


Fig. 8. Distribution of locations of persons (upper panels) and response category thresholds (lower panels) with superimposed uncertainties (95% CIs, right y-axes) of the observed person locations along the measurement range (x-axes). Panel (a) represents the six activity items and panel (b) the seven impairment items.

These observations (Table 5, Fig. 8) are likely related to the multidimensionality identified with the full UPDRS II scale as well as to the local response dependence that is suggested by relatively large negative item fit residual values among the activity items (Table 5). That is, it has been shown that multidimensionality (as with the original 13-item UPDRS II) tends to decrease reliability and dispersion of person locations, whereas the opposite is seen with local response dependence [44]. The relatively narrow locations of response category thresholds in the impairment scale (Fig. 8b, lower panel) also impacts the precision of person estimates. In addition, narrow response category thresholds imply compromised measurement precision, i.e., that impairment items are more difficult to assess. The measurement uncertainty of the impairment scale (Fig. 8b, upper panel) at the best point of measurement (around 0 logits) is similar to that of the original UPDRS II (95% CIs of ±0.66 vs. ±0.55) but increases rapidly as one moves down the latent continuum. For the activity scale, which has a larger spread of response category thresholds, the corresponding measurement uncertainty (Fig. 8a, upper panel) is larger compared to the UPDRS II (95% CIs of ±1.12 vs. ±0.55) but remains relatively stable over a fairly wide range of the latent continuum. However, it should be kept in mind that since the frames of reference for the two item sets are different, it is also likely that the unit of measurement differs between the two [80]. Therefore, the two are not directly comparable with respect to their ranges of measurement and related aspects.

As noted above, the pattern of primarily relatively large negative fit residuals among activity items suggest the presence of local response dependence among items 5-7. Items 4 and 8 on the other hand tend to show signs of potential multidimensionality (Table 5). However, the chi-square statistics of the activity items do not suggest any major problems and inspection of the associated ICCs suggest that departures from model expectations are benign (Fig. 9a and b). This is also the case for impairment items, although the somewhat curious pattern of item 12 (Fig. 5b) persisted when these items were analysed as a separate scale (data not shown).

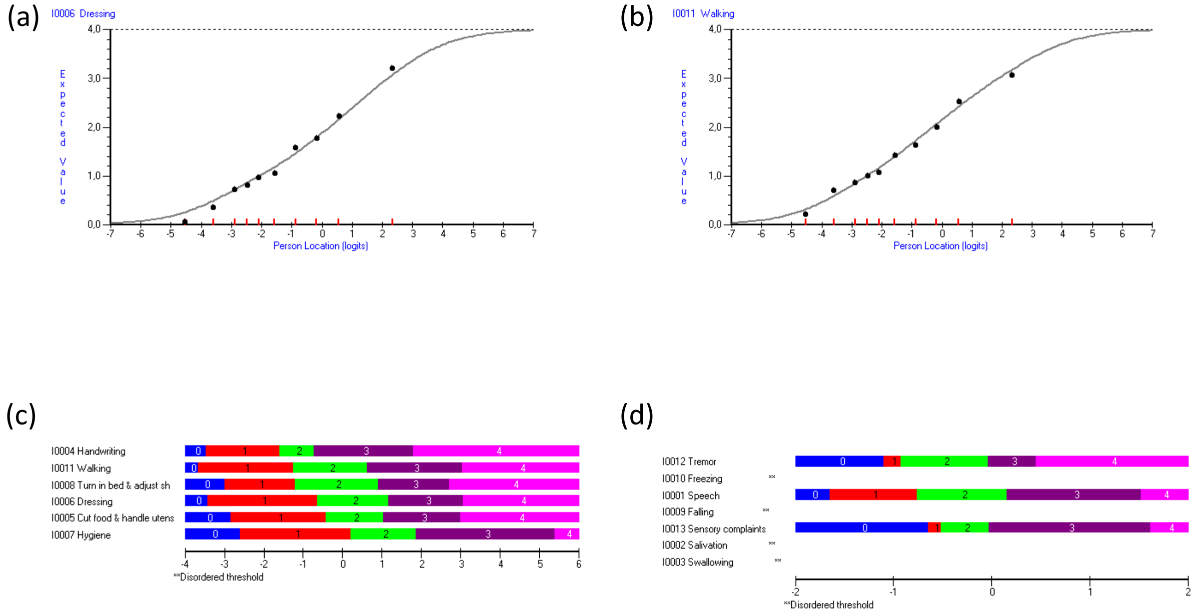


Fig. 9. Panels (a) and (b) displays model fit graphically for the worst (item 6) and best (item 11) fitting activity items, respectively. Panels (c) and (d) provide overviews of response category functioning for activity and impairment items, respectively, as ordered hierarchically from the top.

There was no DIF associated with the activity scale, whereas the DIF identified with five of the seven impairment items in the original UPDRS II remained (Table 5). Similarly, whereas the response categories for activity items worked as expected (Fig. 9c), the four impairment items with disordered response category thresholds in the initial analysis remained disordered (Fig. 9d) according to very similar patterns as those seen earlier (Fig. 6).

Figure 9 also illustrates the empirical hierarchies among activity (Fig. 9c) and impairment (Fig. 9d) items. Again, these are illustrated together with their respective location uncertainties in Figure 7. Since this appears to be the first investigation of the UPDRS II using RMT, it is not possible to relate these results to previous data. However, RMT was used to address the measurement properties of the patient-reported PDQ-39 [81], which include a 6-item ADL scale that has four items in common with the UPDRS II. Interestingly, the relative item ordering among the four UPDRS II items (from lower to higher locations: items 4 – 6 – 5 – 7; Table 5) was very similar to that of their PDQ-39 counterparts (writing – cutting food – dressing – washing) [81]. In fact, when accouting for the overlap in location 95% CIs, the respective hierarchies are indistinguishable. Given the lack of a conceptual theory for the UPDRS II as well as for the PDQ-39 [15, 74], this provides some very tentative evidence of a partial structure for activity performance in people with PD.

Taken together, observations from post hoc analyses of the UPDRS II divided into two conceptually derived item sets representing activities and impairments, respectively, suggest that a reasonable albeit suboptimal measure of ADL outcomes may be derived from the 6-item activity scale. However, the 7-item impairment scale appears unable to contribute towards measurement as evidenced by, e.g., an overly restricted range of measurement and inability to separate the sample into even two distinct groups of people. These items are therefore likely to contribute little but noise to the total UPDRS II score and should be omitted from any use of the UPDRS II that attempts to assess or measure ADL in PD. In their current format, these impairment items primarily appear to represent a clinical check-list rather than being part of a measurement system. As such, responses to these individual items may provide clinically valueable information by their own virtue, but this should not be confused with measurement.

5.5. On a road to nowhere?

As noted in section 1.2, a revised version of the UPDRS has been proposed [18, 19]. However, a review of items in part II of the MDS-UPDRS does not suggest that the revised version would exhibit any substantial benefits as compared to the original UPDRS II. That is, about half of the 13 MDS-UPDRS II items still represent impairments rather than activities [16]. Furthermore, testing and development of the scale still relied on CTT (primarily factor analyses), and aspects of the final results appear to have been misinterpreted within that frame of reference [19]. Indeed, recent RMT based analyses of the MDS-UPDRS have failed to show any clear advantages over the original version presented here [82]. With this in mind, it is interesting to review early attemps to evaluate ADL outcomes, such as that suggested by Duff [11] some 60 years before the MDS-UPDRS (see section 1.2). Considering the Duff scale from an ICF perspective as was done with parts II of the UPDRS and MDS-UPDRS reveals that at least 9 of its 10 items indeed appear to represent activities [16]. However, further and more thorough conceptual reviews integrated with careful RMT analyses will be need to allow for firm conclusions.

Another more recent development is the Penn Parkinson's Daily Activities Questionnaire (PDAQ) [83]. However, the PDAQ differs from the UPDRS II as well as other ADL instruments in some central respects. First, it is not based on patient-report or clinical observation but is intended to be completed by family members or others with daily contact with the patient. Second, its items lack many central ADL aspects (e.g., those related to mobility) but emphasises cognitive aspects (e.g., attention, comprehension, learning and memory) of activity performance. Furthermore, the PDAQ was developed based on item response theory (IRT), a family of psychometric models that share certain features with RMT but differs in some key aspects. Essentially, whereas RMT articulates requirements for measurement that data should meet, IRT is a data modelling approach where item parameters additional to the location are introduced to account for the data and the model that best fits the data is sought [84-86]. In effect, the introduction of additional item parameters counteracts central measurement properties such as invariance and sufficiency, which has caused IRT to be challenged as a valid approach to measurement [87].

While RMT is increasingly applied in health outcome measurement, CTT remains the dominant paradigm. In view of this and the contrasting results between CTT and RMT presented here it is interesting to consider the results from a recent study by Maul, in which empirical responses to items experimentally designed to measure nothing were analyzed using common CTT methodologies such as coefficient alpha and factor analysis [88]. It was found that items inquiring about either a phenomenon that does not exist ("gavagai"), nonsense words (stock lorem ipsum text) or completely blank items (response categories without items) rendered responses that met common CTT criteria and did not necessarily perform less well than real items did. Indeed, comparing those results with the CTT based analyses presented here does not suggest that the UPDRS II appears convincingly better than, e.g., meaningless lorem ipsum items.

While Maul [88] did not test his experimental items using the Rasch model, it must be emphasized that since this model (like any numerically based analysis) in itself only processes numbers, there is no guarantee that it would have yielded results contradicting those based on CTT. For example, Hobart et al. [89] tested an instrument intended to measure fatigue that is commonly used in clinical trials and other reseach. They found that it appeared to work well according to CTT as well as RMT based analyses, despite the fact that a multidisciplinary Delphi panel of experienced health care professionals agreed that none of its items was fatigue specific [89].

These experiences illustrate the fundamental importance of a sound theoretical and conceptual basis in order to achieve meaningful measurement. While this is true regardless of what methodology is used for measurement quality assurance, it is inherently embedded and integrated in RMT [30, 39, 76, 90]. In addition, RMT represents an approach to latent variable measurement that in contrast to CTT is based on measurement principles from the physical sciences and has been considered to belong to the same class of models that metrologists consider paradigmatic of measurement [39, 91-93]. This has also been supported by empirical demonstrations that RMT (in contrast to, e.g., CTT and IRT) is the psychometric approach that best enables measurement [87]. It is therefore encouraging to see that the application of RMT is proliferating in the health sciences. However, to fully benefit from its potentials and move towards true advancement within health outcome measurement it is suggested that the field should focus on coordinating efforts towards conceptually based variable definitions and integrations into quality assured measurement systems [76, 94, 95].

6. Conclusions

This paper has demonstrated the central role of health outcome measures in clinical research, and their impact on decisions regarding the value of therapies, as exemplified by reviewing an RCT aiming to demonstrate ADL benefits in people with PD by using the UPDRS II as the primary outcome measure. Careful examination of the UPDRS II through reviewing the literature, conceptual considerations and analyses of empirical data were conducted to better understand RCT results and the role of the UPDRS II as an outcome measure. CTT analyses corroborated previous studies by suggesting that the UPDRS II is adequate as an ADL outcome measure. However, conceptual considerations challenge this conclusion. Application of RMT to the same data revealed a fundamentally different picture that seriously challenge the validity of the UPDRS II as an ADL outcome measure. Splitting the scale into two item sets representing the two identified conceptual constructs of activity limitations and impairments resulted in a potentially useful activity measure, and a collection of impairment items that may or may not provide clinically valueable information.

The experiences from this "case study" appear to be symptomatic for legacy instruments, which still dominate outcome measurement in the clinical sciences. It is therefore time for clinicians and clinical investigators to take advantage of advances in latent variable measurement. This would not only be likely to create new opportunities to improve outcome measurement and patient care, but it may also provide a means to better understand illness.

Acknowledgements

The author wants tho thank the patients, nurses, neurologists and investigators who participated in the original studies that provided data for the analyses for their cooperation, as well as four anonymous reviewers for their constructive comments on the first version of this paper.

Funding: This work was supported by the Swedish Parkinson Foundation, the Strategic Research Area in Neuroscience at Lund University (MultiPark), the Swedish Parkinson Academy, the Swedish Research Council, and Kristianstad University, Kristianstad, Sweden.

References

[1] J. Parkinson, An essay on the shaking palsy, Whittingham and Rowland for Sherwood, Needly and Jones, London, 1817.

[2] L.M. de Lau, M.M. Breteler, Epidemiology of Parkinson's disease, Lancet Neurol, 5 (2006) 525-535.

[3] J. Jankovic, Parkinson's disease: clinical features and diagnosis, J Neurol Neurosurg Psychiatry, 79 (2008) 368-376.

[4] A.H.V. Schapira, K.R. Chaudhuri, P. Jenner, Non-motor features of Parkinson disease, Nature reviews, 18 (2017) 435-450.

[5] A. Haahr, M. Kirkevold, E.O. Hall, K. Ostergaard, Living with advanced Parkinson's disease: a constant struggle with unpredictability, J. Adv. Nurs., 67 (2011) 408-417.

[6] B. Habermann, Day-to-day demands of Parkinson's disease, West. J. Nurs. Res., 18 (1996) 397-413.

[7] S. Hartley, M. McArthur, M. Coenen, M. Cabello, V. Covelli, J. Roszczynska-Michta, T. Pitkanen, J. Bickenbach, A. Cieza, Narratives reflecting the lived experiences of people with brain disorders: common psychosocial difficulties and determinants, PLoS One, 9 (2014) e96890.

[8] S. Przedborski, The two-century journey of Parkinson disease research, Nature reviews, 18 (2017) 251-259.

[9] E. Tolosa, M.J. Marti, F. Valldeoriola, J.L. Molinuevo, History of levodopa and dopamine agonists in Parkinson's disease treatment, Neurology, 50 (1998) S2-10; discussion S44-18.

[10] S.J. Cano, J.C. Hobart, The problem with health measurement, Patient preference and adherence, 5 (2011) 279-290.

[11] R.S. Duff, Use of diparcol in parkinsonism, Br. Med. J., 1 (1949) 613-615.

[12] A.C. England, Jr., R.S. Schwab, Postoperative medical evaluation of 26 selected patients with Parkinson's disease, J. Am. Geriatr. Soc., 4 (1956) 1219-1232.

[13] G.J. Canter, R. De La Torre, M. Mier, A method for evaluating disability in patients with Parkinson's disease, J. Nerv. Ment. Dis., 133 (1961) 143-147.

[14] G.C. Cotzias, M.H. Van Woert, L.M. Schiffer, Aromatic amino acids and modification of parkinsonism, N. Engl. J. Med., 276 (1967) 374-379.

[15] S. Fahn, R.L. Elton, members of the UPDRS development committee, Unified Parkinson's Disease Rating Scale, in: S. Fahn, C.D. Marsden, D.B. Calne, M. Goldstein (Eds.) Recent Developments in Parkinson's Disease, Vol. 2, MacMillan Healthcare Information, Florham Park, 1987, pp. 153-163.

[16] World Health Organization., International classification of functioning, disability and health: ICF, World Health Organization, Geneva, 2001.

[17] S.G. Diamond, C.H. Markham, Evaluating the evaluations: or how to weigh the scales of parkinsonian disability, Neurology, 33 (1983) 1098-1099.

[18] C.G. Goetz, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, G.T. Stebbins, M.B. Stern, B.C. Tilley, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A.E. Lang, A. Lees, S. Leurgans, P.A. LeWitt, D. Nyenhuis, C.W. Olanow, O. Rascol, A. Schrag, J.A. Teresi, J.J. Van Hilten, N. LaPelle, Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan, Mov Disord, 22 (2007) 41-47.

[19] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M.B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A.E. Lang, A. Lees, S. Leurgans, P.A. LeWitt, D. Nyenhuis, C.W. Olanow, O. Rascol, A. Schrag, J.A. Teresi, J.J. van Hilten, N. LaPelle, U.R.T.F. Movement Disorder Society, Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease

Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results, Mov Disord, 23 (2008) 2129-2170.
[20] R.E. Traub, Classical test theory in historical perspective, Educational Measurement: Issues and Practice, 16 (1997) 8-14.
[21] J.C. Nunnally, I.H. Bernstein, Psychometric theory, 3rd ed., McGraw-Hill, Inc., New York, 1994.
[22] F. Galton, Psychometric experiments, Brain, 2 (1879) 149-162.
[23] M. Wilson, Using the concept of a measurement system to characterize measurement models used in psychometrics, Measurement, 46 (2013) 3766-3774.
[24] L.J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika, 16 (1951) 297-334.
[25] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and psychological measurement, 33 (1973) 613-619.
[26] D.L. Streiner, G.R. Norman, Health measurement scales: a practical guide to their development and use, 4th ed., Oxford University Press Inc., New York, 2008.
[27] L.L. Thurstone, The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems, Edwards Brothers, Inc., Ann Arbor, MI, 1931.
[28] J.B. Carroll, The nature of the data, or how to choose a correlation coefficient, Psychometrika, 26 (1961) 347-372.
[29] E. Svensson, Guidelines to statistical evaluation of data from rating scales and questionnaires, J Rehabil Med, 33 (2001) 47-48.
[30] G. Rasch, Probabilistic models for some intelligence and attainment tests, Danmarks Paedagogiske Institut, Copenhagen, 1960.
[31] D. Andrich, B. Sheridan, G. Luo, Interpreting RUMM2030, RUMM Laboratory Pty Ltd., Perth, 2013.
[32] D. Andrich, A rating formulation for ordered response categories, Psychometrika, 43 (1978) 561-573.
[33] D. Andrich, Rating scales and Rasch measurement, Expert Rev Pharmacoecon Outcomes Res, 11 (2011) 571-585.
[34] G. Luo, The relationship between the Rating Scale and Partial Credit Models and the implication of disordered thresholds of the Rasch models for polytomous responses, J Appl Meas, 6 (2005) 443-455.
[35] B.D. Wright, G.N. Masters, Rating scale analysis, MESA Press, Chicago, 1982.
[36] G.N. Masters, A Rasch model for partial credit scoring, Psychometrika, 47 (1982) 149-174.
[37] D. Andrich, An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any “threshold disorder controversy”, Educational and Psychological Measurement, 73 (2013) 78-124.
[38] R.J. Adams, M.L. Wu, M. Wilson, The Rasch rating model and the disordered threshold controversy, Educational and Psychological Measurement, 72 (2012) 547-573.
[39] D. Andrich, Rasch models for measurement, Sage Publications, Inc., Beverly Hills, 1988.
[40] G. Rasch, On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements, Danish Yearbook of Philosophy, 14 (1977) 58-93.
[41] R.M. Smith, Fit analysis in latent trait measurement models, J Appl Meas, 1 (2000) 199-218.
[42] R.M. Smith, C. Plackner, The family approach to assessing fit in Rasch measurement, J Appl Meas, 10 (2009) 424-437.

[43] K.B. Christensen, S. Kreiner, Item fit statistics, in: K.B. Christensen, S. Kreiner, M. Mesbah (Eds.) Rasch models in health, John Wiley & Sons, Inc., Croydon, Surrey, 2013, pp. 83-103.
[44] I. Marais, D. Andrich, Formalizing dimension and response violations of local independence in the unidimensional Rasch model, J Appl Meas, 9 (2008) 200-215.
[45] I. Marais, Local dependence, in: K.B. Christensen, S. Kreiner, M. Mesbah (Eds.) Rasch models in health, John Wiley & Sons, Inc., Croydon, Surrey, 2013, pp. 111-130.
[46] J.A. Teresi, Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics, Med. Care, (2006) S152-S170.
[47] C. Hagquist, D. Andrich, Recent advances in analysis of differential item functioning in health research using the Rasch model, Health Qual Life Outcomes, 15 (2017) 181.
[48] J. Brodersen, D. Meads, S. Kreiner, H. Thorsen, L. Doward, S. McKenna, Methodological aspects of differential item functioning in the Rasch model, Journal of medical economics, 10 (2007) 309-324.
[49] J. Hobart, S. Cano, Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods, Health Technol. Assess., 13 (2009) iii, ix-x, 1-177.
[50] P. Hagell, A. Westergren, Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (RUMM) Program in Health Outcome Measurement, J Appl Meas, 17 (2016) 416-431.
[51] C.A. McHorney, P.O. Monahan, Applications of Rasch analysis in health care, Med Care 42 (1 Suppl), 2004, I73-78.
[52] T. Salzberger, The validity of polytomous items in the Rasch model-The role of statistical evidence of the threshold order, Psychological Test and Assessment Modeling 57, 2015, 377-395.
[53] J. Hobart, Rating scales for neurologists, J Neurol Neurosurg Psychiatry, 74 Suppl 4 (2003) iv22-iv26.
[54] K.H. Deane, H. Flaherty, D.J. Daley, R. Pascoe, B. Penhale, C.E. Clarke, C. Sackley, S. Storey, Priority setting partnership to identify the top 10 research priorities for the management of Parkinson's disease, BMJ open, 4 (2014) e006434.
[55] A.N. Nisenzon, M.E. Robinson, D. Bowers, E. Banou, I. Malaty, M.S. Okun, Measurement of patient-centered outcomes in Parkinson's disease: What do patients really want from their treatment?, Parkinsonism Relat Disord, 17 (2011) 89-94.
[56] Food and Drug Administration, Patient-Reported Outcome measures: Use in Medicinal Product Development to Support Labelling Claims, Fed. Regist., 71 (2006) 5862-5863.
[57] Food and Drug Administration, Patient-Reported Outcome measures: Use in Medicinal Product Development to Support Labelling Claims, Washington DC, 2009.
[58] Food and Drug Administration, Clinical Outcome Assessment Qualification Program. Available from: www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm (accessed September 21, 2018).
[59] C.G. Goetz, W. Poewe, O. Rascol, C. Sampaio, G.T. Stebbins, S. Fahn, A.E. Lang, P. Martinez-Martin, B. Tilley, B. van Hilten, C. Kleczka, L. Seidl, The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations, Mov Disord, 18 (2003) 738-750.
[60] L.M. Shulman, M. Armstrong, T. Ellis, A. Gruber-Baldini, F. Horak, A. Nieuwboer, S. Parashos, B. Post, M. Rogers, A. Siderowf, Disability rating scales in Parkinson's disease: critique and recommendations, Mov. Disord., 31 (2016) 1455-1465.
[61] M.M. Hoehn, M.D. Yahr, Parkinsonism: onset, progression and mortality, Neurology, 17 (1967) 427-442.

[62] J.E. Ware, Jr., B. Gandek, Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach. International Quality of Life Assessment, J Clin Epidemiol, 51 (1998) 945-952.
[63] J. Baglin, Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR, Practical Assessment, Research & Evaluation, 19 (2014) 2.
[64] A.M. Gadermann, M. Guhn, B.D. Zumbo, Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide, Practical Assessment, Research & Evaluation, 17 (2012).
[65] M.E. Timmerman, U. Lorenzo-Seva, Dimensionality assessment of ordered polytomous items with parallel analysis, Psychol Methods, 16 (2011) 209-220.
[66] A.B. Costello, J. Osborne, Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis, Practical Assessment Research & Evaluation, 10 (2005).
[67] U. Lorenzo-Seva, P.J. Ferrando, FACTOR: a computer program to fit the exploratory factor analysis model, Behavior research methods, 38 (2006) 88-91.
[68] J.M. Bland, D.G. Altman, Multiple significance tests: the Bonferroni method, BMJ, 310 (1995) 170.
[69] A. Schrag, C. Sampaio, N. Counsell, W. Poewe, Minimal clinically important change on the unified Parkinson's disease rating scale, Mov Disord, 21 (2006) 1200-1207.
[70] P. Martinez-Martin, L. Prieto, M.J. Forjaz, Longitudinal metric properties of disability rating scales for Parkinson's disease, Value Health, 9 (2006) 386-393.
[71] O. Rascol, Defining a minimal clinically relevant difference for the unified Parkinson's rating scale: an important but still unmet need, Mov Disord, 21 (2006) 1059-1061.
[72] T. Steffen, M. Seney, Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism, Phys. Ther., 88 (2008) 733-746.
[73] G.M. Hariz, M. Lindberg, M.I. Hariz, A.T. Bergenheim, Does the ADL part of the unified Parkinson's disease rating scale measure ADL? An evaluation in patients after pallidotomy and thalamic deep brain stimulation, Mov Disord, 18 (2003) 373-381.
[74] V. Peto, C. Jenkinson, R. Fitzpatrick, R. Greenhall, The development and validation of a short measure of functioning and well being for individuals with Parkinson's disease, Qual Life Res, 4 (1995) 241-248.
[75] T.S. Kuhn, The function of measurement in modern physical science, Isis, 52 (1961) 161-193.
[76] A.J. Stenner, W.P. Fisher, Jr., M.H. Stone, D.S. Burdick, Causal Rasch models, Frontiers in psychology, 4 (2013) 536.
[77] M.T. King, A point of minimal important difference (MID): a critique of terminology and methods, Expert review of pharmacoeconomics & outcomes research, 11 (2011) 171-184.
[78] D. Revicki, R.D. Hays, D. Cella, J. Sloan, Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes, J Clin Epidemiol, 61 (2008) 102-109.
[79] D. Andrich, An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern, Education Research & Perspectives, 9 (1982) 95-104.
[80] S.M. Humphry, D. Andrich, Understanding the unit in the Rasch model, J Appl Meas, 9 (2008) 249-264.
[81] P. Hagell, C. Nygren, The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine, J Neurol Neurosurg Psychiatry, 78 (2007) 1191-1198.

[82] A. Regnault, B. Boroojerdi, J. Meunier, M. Bani, S. Cano, A psychometric analysis of the UPDRS-II and -III in early Parkinson's disease: Do the numbers add up?, The 22nd International Congress of Parkinson's Disease and Movement Disorders, Hong Kong, 2018.
[83] L. Brennan, A. Siderowf, J.D. Rubright, J. Rick, N. Dahodwala, J.E. Duda, H. Hurtig, M. Stern, S.X. Xie, L. Rennert, Development and initial testing of the Penn Parkinson's Daily Activities Questionnaire, Mov. Disord., 31 (2016) 126-134.
[84] D. Andrich, Controversy and the Rasch model: a characteristic of incompatible paradigms?, Med Care, 42 (1 Suppl), 2004, I7-16.
[85] R.W. Massof, Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires, Ophthalmic Epidemiol., 18 (2011) 1-19.
[86] J. Petrillo, S.J. Cano, L.D. McLeod, C.D. Coon, Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples, Value in Health, 18 (2015) 25-34.
[87] R.W. Massof, The measurement of vision disability, Optom. Vis. Sci., 79 (2002) 516-552.
[88] A. Maul, Rethinking traditional methods of survey validation, Measurement: Interdisciplinary Research and Perspectives, 15 (2017) 51-69.
[89] J. Hobart, S. Cano, R. Baron, A. Thompson, S. Schwid, J. Zajicek, D. Andrich, Achieving valid patient-reported outcomes measurement: a lesson from fatigue in multiple sclerosis, Mult. Scler., 19 (2013) 1773-1783.
[90] M. Wilson, Constructing measures: an item response modelling approach, Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2005.
[91] L. Mari, M. Wilson, An introduction to the Rasch measurement approach for metrologists, Measurement, 51 (2014) 315-327.
[92] L. Pendrill, Man as a Measurement Instrument, NCSLI Measure, 9 (2014) 24-35.
[93] W.P. Fisher Jr, Invariance and traceability for measures of human, social, and natural capital: Theory and application, Measurement, 42 (2009) 1278-1287.
[94] S.J. Cano, L. Pendrill, S.P. Barbic, W.P. Fisher Jr, Patient-centred outcome metrology for healthcare decision-making, Journal of Physics: Conference Series, IOP Publishing, 2018, pp. 012057.
[95] W.P. Fisher, Invariance and traceability for measures of human, social, and natural capital: Theory and application, Measurement, 42 (2009) 1278-1287.

Figure Legends

Figure 1.

Illustration of the basic instrument design and assumptions underpinning latent variable measurement. Items are observable manifestations of the unobservable latent target variable and are expected to reflect variations in the latent variable. Observed item responses form the basis in the measurement process used to locate the individual on a latent quantitative continuum intended to represent her/his position on the target variable, fromm less to more.

Figure 2.

Schematic illustration of the design and results of a randomized controlled phase IV trial of an adjunct active drug (A) or placebo (P) used together with levodopa (LD) in people with fluctuating Parkinson's disease (PD) to determine the effect on activities of daily living (ADL) using the UPDRS II as the primary outcome measure.

Figure 3.

Scree plots of the eigenvalues (y axes) for components (x-axis, panel a) and factors (x-axis, panel b) identified by principal component analysis (PCA; panel a) and minimum rank factor analysis (MRFA; panel b) of item-level UPDRS II data. The dashed horizontal line indicates the cut-off point for determination of the number of components and factors according to the eigenvalue >1 criterion.

Figure 4.

(a) Distribution of locations of persons (upper panel) and UPDRS II response category thresholds (lower panel) on the common logit metric (*x*-axis; negative values = better). Superimposed black dot represent the uncertainties (95% CIs, right y-axis) of the observed person locations along the measurement range. (b) Relationship between raw total UPDRS II scores (y-axis) and their implied linear locations on the logit metric (x-axis) with estimated lower and upper 95% CI limits of uncertainty (±1.96*SE*; represented by horizontal error bars) across the full range of all possible UPDRS II raw total scores.

Figure 5.

Graphical representation of tests of fit between UPDRS II items and the Rasch model. Panel (a) Item chi-square statistics plotted in ascending order with numbers representing item numbers. Panels (b) through (d) display item characteristic curves (ICCs) of expected (grey curves) and observed (black dots) item responses (y-axis) for the 10 class intervals (subgroups of people with different locations; n=41-59 per class interval) along the outcome continuum (x-axis). Panels (b) and (c) display display the worst fitting items (items 12 and 13, respectively). For comparison, panel (d) represents an item with relatively good fit (item 4).

Figure 6.

Response category functioning for UPDRS II items. Panel (a) provides an overview of all items ordered hierarchically from the top. Intersections between color bars represent response category thresholds. Instances of disordered thresholds are blank since disordering prevents this type of graphical representaation. Panels (b) through (d) provide a more detailed picture, where each colored category probability curve represenst the probability (y-axis) of responding in that response category relative to various estimated person logit locations (x-axis). Panels (b) and (c) illustrate patterns of disordered thresholds for items 2 and 9, respectively. For comparison, panel (d) illustrates an item without disordered thresholds (item 8).

Figure 7.
Hierarchical item ordering with item locations on the y-axis (negative values = better) and item numbers on the x-axis (see Table 1 for item descriptions). Error bars are uncertainties (±95% CIs). Panel (a) represents the total UPDRS II score, whereas panels (b) and (c) displays the item hierarchies when analyzing activity (b) and impairment (c) items separately.

Figure 8.
Distribution of locations of persons (upper panels) and response category thresholds (lower panels) with superimposed uncertainties (95% CIs, right y-axes) of the observed person locations along the measurement range (x-axes). Panel (a) represents the six activity items and panel (b) the seven impairment items.

Figure 9.
Panels (a) and (b) displays model fit graphically for the worst (item 6) and best (item 11) fitting activity items, respectively. Panels (c) and (d) provide overviews of response category functioning for activity and impairment items, respectively, as ordered hierarchically from the top.