Journal of Huntington's Disease 13 (2024) 385–397
DOI 10.3233/JHD-240001
IOS Press

Research Report

Rasch Measurement Theory (RMT) Analyses of the Huntington's Disease Everyday Functioning (Hi-DEF) to Evaluate Item Fit and Performance

Jennifer Petrillo[a,*], Ruta Sawant[a], Emma Elliott[b], Sophie Cleanthous[b], Rebecca Rogers[b], Stefan Cano[b], Sarah Baradaran[a] and Jason Johannesen[a]
[a]*Sage Therapeutics, Inc., Cambridge, MA, USA*
[b]*Modus Outcomes, a Division of THREAD, St. James Square, Cheltenham, England, UK*

Accepted 28 May 2024
Pre-press 23 August 2024
Published 10 September 2024

Abstract.
Background: The Huntington's Disease (HD) Everyday Functioning (Hi-DEF) is a new patient-reported outcome (PRO) instrument designed to measure the impact of cognitive impairment on daily functioning in the early stages of HD.
Objective: To assess the measurement properties and finalize item content of the Hi-DEF.
Methods: A cross-sectional, observational psychometric validation study was conducted among individuals with early stages of HD at 9 US centers of excellence. Rasch Measurement Theory (RMT) analysis of the initial draft version of the Hi-DEF (47 items) and subscales (i.e., 'Home', 'At work', 'Driving', and 'Communication') was conducted to examine measurement properties including sample-to-scale targeting, suitability of response scale (ordering of response thresholds), scale cohesiveness (item fit), local independence, and person fit.
Results: 151 participants (mean age 47 years (SD 12), 59% female) were included. Seven items were removed based on dependency and item fit. The remaining 40-item version of the Hi-DEF demonstrated good measurement properties. Across the four subscales, targeting ranged from 49–70% (72% full scale), reliability ascertained by person separation index ranged from 0.53–0.87 (0.92 full scale), response scales were ordered for 25–100% of items (75% full scale), 0–12% items displayed misfit (2% full scale), and 0–1% (2% full scale) item pairs displayed dependency.
Conclusions: Our study supports the psychometric integrity of the Hi-DEF as a reliable and valid new PRO instrument designed to assess the impact of cognitive impairment on daily functioning in the early stages of HD. Future work will evaluate the external validity and utility in clinical trial applications.

Keywords: Huntington's disease, executive function, activities of daily living, psychometrics, Hi-DEF, patient-reported outcomes

INTRODUCTION

Huntington's disease (HD) is a progressive neurodegenerative disease caused by an autosomal dominantly inherited CAG trinucleotide repeat

*Correspondence to: Jennifer Petrillo, Sage Therapeutics, Inc., 55 Cambridge Parkway, Cambridge, MA 02142, USA. Tel.: +1 617 299 8380; E-mail: jennifer.billet@sagerx.com.

expansion in the huntingtin (HTT) gene, which results in the production of mutant HTT protein [1]. Around 40,000 people in the US currently have manifest HD [2], with occurrence rates of 10.6 to 13.7 per 100,000 in Western populations and a lower incidence (1 to 7 per million) in Eastern regions and countries.

HD is characterized by a triad of motor, cognitive and psychiatric symptoms. Typically, cognitive symptom onset begins in adults around 40 years old (range 30–59 years), with signs and symptoms developing gradually over the course of one or two decades [3, 4]. While HD diagnosis is primarily based on motor symptom manifestation, cognitive and psychiatric symptoms develop much earlier (up to 15 years prior to motor manifestations) and impose significant burden on patients and families [5–9]. Researchers have mounted a fundamental challenge to the prevailing clinical classification of HD, which relies on the presence of involuntary movements and genetic tests, in favor of the evaluation of neuropsychiatric symptoms and signs of cognitive impairment [9]. Research suggests that significant decline in executive function can be detected in early stages of HD, involving cognitive processes such as working memory, planning, organization, initiation, cognitive flexibility, decision making, problem solving, selective attention, and inhibitory control [10–13] as well as visuospatial performance.

Cognitive impairment in HD has a profound effect on an individual's daily living and quality of life, including their ability to function at home, to drive and to communicate with others [1, 14]. As cognitive functioning deteriorates, overall functioning is impacted, often necessitating job modification or even loss of employment [1]. Collectively, this impairment can lead to increased healthcare resource use and costs. It is therefore important to identify cognitive impairment early in the disease course, to allow for timely intervention and implementation of treatments including coping strategies.

Cognitive tests conducted using performance outcome (PerfO) instruments are typically used to assess cognition in HD and are considered a more objective measure of cognitive ability [15]. However, there is increasing interest in patient-reported outcome (PRO) instruments, which allow an understanding of an individual's own experiences of their abilities and the perceived impact of any cognitive impairment on their day-to-day functioning. PROs and quality of life measures such as the HD-PRO-TRIAD™, Huntington Disease Health-Related Quality of Life (HDQLIFE™), and Functional Rating Scale 2.0 (FurST 2.0) are valid and reliable measures of HD-specific health-related functioning and quality of life [16–18]. While these scales assess the broader aspects of cognitive impairment, they may not be optimized for sensitivity to subtle cognitive changes that begin to occur early in the course of HD. Although early cognitive changes may not severely limit activities of daily living, reductions in higher order executive abilities can affect more complex aspects of work and home life, driving, and social functioning.

The Hi-DEF, a novel PRO instrument, was developed to address this gap based on findings from preliminary qualitative research that has been described in detail elsewhere [14]. This research provided support for the content validity of the Hi-DEF and that the instrument and its concepts are clear, relevant, accepted, and interpreted as intended by individuals with HD. Furthermore, they revealed the impact of executive functioning impairments in the early stages of HD and the associated decline in daily functioning. A draft measure was developed containing 47 items over four subscales measuring the impact of cognitive impairment at home, at work, and while driving and communicating. This paper describes the initial psychometric evaluation of the Hi-DEF conducted using data from an observational validation study using Rasch Measurement Theory (RMT) analyses (a recommended approach for the development of novel PROs [19–21]), to inform the finalization of the item content, the scoring structure, as well as an assessment of the overall measurement properties. Further psychometric analyses of the Hi-DEF on the basis of this observational study, including assessment of construct validity with other clinical outcome assessments will be presented in a separate paper. Preliminary results of this validation have been published elsewhere [22].

METHODS

Study population and sampling

An observational validation study was conducted virtually at 9 HD centers of excellence (COEs) across the US. The COEs helped in recruitment and assessment of participants for this study.

Inclusion criteria for participants included: (i) ability to read and respond to questionnaire items in English, (ii) age range of 25 to 65 years of age, (iii) huntingtin (HTT) mutation gene positive carrier,

and (iv) investigator-confirmed cognitive changes or Total Functional Capacity (TFC) score ≥8 [23]. The exclusion criteria included an ongoing neurological condition other than HD that, in the opinion of the investigator, may have impacted the participants' current cognitive or motor symptoms, or had a history of neurosurgical intervention or significant head injury. Participants received compensation of $150 for participation. This study received ethical approval from Advarra IRB (ref# Pro00048743).

Participants completed a demographic and medical history form, two PRO instruments, the Hi-DEF scale and HD-PRO-Triad (measuring cognitive, behavioral, and motor symptoms in HD) [16], and a battery of cognitive assessments (Spatial Span, One Touch Stockings of Cambridge, Spatial Working Memory, Emotion Recognition Task, and Paired Associate Learning). All of the above instruments were administered to participants on the Cambridge Neuropsychological Test Automated Battery (CANTAB) Connect platform, a cognitive assessment research software, via the web on a laptop or computer screen either at their home or during an onsite visit.

The study visit was completed either at the clinical site or at the participant's home. The visit consisted of PRO and demographic questionnaire completion (20 min), a practice cognitive assessment session (30 min), a mandatory break (60 min), and then the scored cognitive assessment session (30 min). Participant medical history, including date of genetic testing, CAG repeats, and Unified Huntington's Disease Rating Scale (UHDRS) TFC score were entered into the platform by study coordinators.

Hi-DEF description

The Hi-DEF is a PRO instrument developed to measure the impact of cognitive impairment on daily living in early-stage HD. The draft 47 items concerned daily life difficulties related to cognitive functioning challenges at 'Home' (17 items), 'At work' (13 items), and while 'Driving' (9 items), all of which were scored on a 5-level difficulty scale (1 = No difficulty, 2 = A little difficulty, 3 = Some difficulty, 4 = A lot of difficulty, 5 = Cannot do this anymore, N/A = Didn't have the opportunity to do this in the past week). In addition, cognitive functioning challenges related to 'Communicating' (8 items) were scored on a 4-level frequency scale (1 = Never, 2 = Sometimes, 3 = Often, 4 = Almost always, N/A = Didn't have the opportunity to do in the past week). The items were completed online and required a response on each item, resulting in no true missing data. Hi-DEF responses 1–5 were rescored as 0–4 for analysis; 'N/A' responses were recorded as missing.

Psychometric analysis method: Rasch measurement theory

Rasch measurement theory (RMT) analyses were used to examine the extent to which observed raw scores on the Hi-DEF meet the scores expected by the Rasch model and subsequently indicating the extent to which the summing of scale items results in rigorous measurement [24–26]. RMT analysis has three broad aims relating to evaluating the extent to which the sample-to-scale targeting is adequate, the measurement continuum has been constructed successfully, and the sample was valid. To this effect, seven RMT-based psychometric properties were examined (Table 1), results of which were interpreted with reference to published guidelines wherever possible [27].

To facilitate interpretation, measures (scores) were transformed into an accessible and intuitively meaningful metric, ranging from 0 (no difficulty) to 100 (maximum difficulty).

RMT analysis process

RMT analysis was performed using RUMM2030 software [28] in two stages. At the first stage, a comprehensive psychometric analysis of the measurement performance of the draft 47-item Hi-DEF scale was performed in line with methods [29–31] described above and in Table 1. Results were reviewed, and any necessary revisions were made to the item content and/or response scale. The revised item content and response scales, comprising the final Hi-DEF, were evaluated using the same RMT methods. Additional construct validity analyses were performed on the final 40-item Hi-DEF using the cognitive assessments and HD-PRO-TRIAD™ administered to patients; those analyses will be described in a separate manuscript.

RESULTS

Demographics

A total of 151 individuals with HD participated in this study. Recruitment was over a period of

Table 1
RMT property definitions and definitions in context of the Hi-DEF scale

RMT property measured	Description	Assessment test and criteria
Targeting & scale coverage	The extent to which the Hi-DEF items measure the full range of HD patients across low to high cognitive impairment/daily functioning difficulties.	There is no specific criterion. Examination involves a visual inspection of the relative distributions of item locations and person measurements on a common scale (graphical indicator) and an estimation of the percentage of participants covered by the scale range, which is expected to be at least 60% for adequate targeting.
Reliability	The extent to which the Hi-DEF items can detect differences in the impact of cognitive impairment on daily functioning within a sample and detect changes over time.	Measured by Person Separation Index (PSI), the PSI ranges from 0 (all error) to 1 (no error). Therefore, the closer to 1.0 the higher the reliability with a minimum of 0.7 reflecting good reliability.
Suitability of response scale (ordering of response item thresholds)	The extent to which response options and scoring functions for Hi-DEF items work as intended to form a continuum from less to more cognitive impairment/daily functioning difficulties in increments that responders can consistently distinguish between.	Examination of the category probability curves (CPCs) which show the ordering of the thresholds for each item. A threshold marks the location on the latent continuum where two adjacent response categories are equally likely. The ordering of the thresholds should reflect the intended order of the categories, i.e., ordered sequentially from less to more.
Item fit*	The extent to which Hi-DEF items work together clinically and statistically, to ensure that items define a cohesive continuum of cognitive impairment/daily functioning difficulties and that it is appropriate to sum single item responses to obtain a total score.	No single "fit" indicator is sufficient. A fit statistic, derived by forming class intervals of participants with similar measurement locations, for which observed scores are expected to be in line with those expected by the Rasch model, are examined statistically using chi-square values with associated probabilities, as well as their graphical counterparts (item characteristic curves; ICCs).
Local independence*	The extent to which Hi-DEF items are dependent upon, or biased by, each other, i.e., whether item responses are locally independent. If responses to one item directly influence responses to another, measurement estimates are artificially inflated or deflated (biased), and reliability is artificially elevated.	Local independence is assessed by investigating the correlations of the item residuals (referred to as residual correlations). As a guide, residual correlations exceeding +/-0.30 warrant further examination as this reflects > 9% of shared variance.
Item stability*	The extent to which the Hi-DEF is stable across different sub-groups (in this case, age and gender) indicating whether the items mean the same to different participant groups.	Examined using analysis of variance (ANOVA) assessing stability of item scores between sample gender and age groups and across the different class-intervals, where a significant result is taken to indicate differential item functioning (DIF).
Person fit*	The extent to which participant response patterns on Hi-DEF items are statistically consistent with the RMT model, indicating whether the measurement is valid.	Examination of person fit residuals which are expected to lie between –2.5 to 2.5 as this indicates participant's response patterns align with the RMT model. A proportion of up to 5% of underfitting persons was considered acceptable.

*Properties informing the assessment of construct validity within RMT, i.e., the extent to which a scale measures what it intended to measure.

12 months, although delays from academic IRB approvals reduced the time available for recruitment for some sites. The majority of participants ($n = 117$; 77.5%) completed the assessment within the expected duration of 3 hours.

The sample was representative of a population in early stages of HD (Table 2): 59% female, with a mean age of 47.3 years (SD 11.6; range 25 to 65), and a mean TFC score of 11.4 (range 8 to 13; score 8, $n = 15$; score 9, $n = 15$; score 10, $n = 13$; score 11, $n = 21$; score 12, $n = 30$; score 13, $n = 55$). A third of participants did not self-report any psychiatric conditions on the demographic and health information form ($n = 50$, 33.1%); however, the majority of participants self-reported at least one psychiatric condition: depression ($n = 72$, 47.7%), anxiety ($n = 79$, 52.3%), mood swings ($n = 26$, 17.2%), obsessive-compulsive symptoms ($n = 16$, 10.6%), and substance or alcohol use disorder ($n = 4$, 2.6%).

The majority of participants were highly educated, with either a bachelor's (34%; $n = 51$) or postgraduate (28%; $n = 42$) degree, and most participants were

Table 2
Sample characteristics

Variable	*n* (missing)		
Age (y)	*n* = 150 (1)	Mean (SD)	47.3 (11.64)
		Range	25 – 65
Sex – *n* (%)	*n* = 150 (1)	Female	89 (59%)
		Male	60 (40%)
		Prefer not to say	1 (<1%)
Ethnicity – *n* (%)	*n* = 134 (17)	Hispanic/Latino	9 (6%)
		Non-Hispanic/Non-Latino	125 (83%)
Race – *n* (%)	*n* = 151 (0)	White	140 (93%)
		Black/African American	8 (5%)
		Multiracial	1 (<1%)
		Other	2 (1%)
Time since genetic test (y)	*n* = 129 (22)	Mean (SD)	5.82 (6.33)
		Minimum, Maximum	0, 27
CAG Repeat	*n* = 142 (9)	Mean (SD)	43.56 (3.30)
		Minimum, Maximum	39.00, 55.00
TFC Score	*n* = 151 (0)	Mean (SD)	11.37 (1.72)
		Minimum, Maximum	8, 13
Education – *n* (%)	*n* = 151 (0)	Some high school	2 (1%)
		High school graduate/GED equivalent	19 (13%)
		Some college	21 (14%)
		Associate degree	11 (7%)
		Bachelor's degree	51 (34%)
		Post graduate degree	42 (28%)
		Trade/technical certification	5 (3%)
Employment status – *n* (%)	*n* = 149 (2)	Working part-time	16 (11%)
		Working full-time	62 (41%)
		Homemaker	10 (7%)
		Retired	15 (9.9%)
		Not employed	9 (6.0%)
		On disability (related to HD)	36 (24%)
		On disability (not related to HD)	1 (<1%)

working full-time (41%; *n* = 62) or part-time (11%; *n* = 16), while 24% (*n* = 36) were on disability related to HD (Table 2). In total, one-fifth (*n* = 30) of participants had a TFC = 13, were working full or part time, and had an associate's, bachelor's, or post-graduate degree. This subgroup of the validation sample provides the opportunity for insight into the performance of the measure at the upper end of functioning. However, further work is needed to assess the broader range of the measure and potential decline over time.

Step 1: RMT analysis of draft 47-items

For full results of the draft instrument, see Table 3. In summary, all subscales with the exception of 'Driving' had high reliability, minimal issues with person fit, and were found to be stable across gender and age groups (i.e., no differential item functioning [DIF]). In terms of sample-to-scale targeting, scales had adequate to good coverage ranging from 61% to 76% for all subscales except for 'Driving', which showed sub-optimal coverage (51%). In terms of item fit, items were largely cohesive, and item dependency was minimal across the four subscales, although there was some dependency in the full scale. Based on the results, 7 items were removed ('Home': 2, 'At work': 2, 'Driving': 1, 'Communicating': 2) to produce the final 40-item Hi-DEF, which consists of the following subscales: 'Home' (15 items), 'At work' (11 items), 'Driving' (8 items), and 'Communicating' (6 items). Supplementary Table 1 summarizes the items removed and corresponding rationale.

Step 2: RMT analysis of the final 40-items

Targeting & scale coverage

Targeting and scale coverage indicates whether the Hi-DEF items measure the full range of HD patients, from low to high cognitive impairment/daily functioning difficulties [32, 33]. A higher coverage rate means that more participants are measured. Sample-to-scale targeting, of the 40-item Hi-DEF

Table 3
Summary of RMT results for the draft 47-item and final 40-item Hi-DEF instrument

		Home		At work		Driving		Communicating		Full scale	
		Draft (17 items)	*Final (15 items)*	*Draft (13 items)*	*Final (11 items)*	*Draft (9 items)*	*Final (8 items)*	*Draft (8 items)*	*Final (6 items)*	*Draft (47 items)*	*Final (40 items)*
Targeting[1]		69%	68%	61%	60%	51%	49%	72%	70%	76%	72%
Response scale[2]		88%	93%	77%	82%	22%	25%	100%	100%	70%	75%
Item fit	Fit residuals[3]	88%	93%	92%	100%	100%	100%	75%	83%	98%	98%
	Chi-square[4]	100%	100%	100%	100%	100%	88%	100%	100%	100%	98%
Item dependency[5]		98%	99%	99%	100%	100%	100%	89%	100%	97%	97%
Reliability (with/without extremes)[6]		0.88/0.91	0.87/0.90	0.86/0.89	0.86/0.89	0.57/0.65	0.53/0.61	0.84/0.82	0.77/0.74	0.93/0.95	0.92/0.94
Item stability (differential item functioning [DIF])		No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex	No DIF for age and sex
Person fit[7]		94%	95%	98%	98%	99%	99%	97%	97%	91%	93%

Higher percentages indicate better findings. [1]Estimated using the percentage of individual sample measurements ($n = 151$) covered by the scale range; [2]Estimated based on the percentage of items displaying ordered response thresholds; [3]Percentage of items with fit residuals inside range of +/–2.5; [4]Percentage of items not displaying significant chi-square estimates; [5]Percentage of item pairs that are not locally dependent based on > 0.3 residual correlations indicating > 9% shared variance; [6]PSI (person separation index) is reported on a scale from 0 to 1 : 0 = all error; 1 = no error; [7]Percentage of persons with fit residuals inside range of +/–2.5.

Table 4
Sample Hi-DEF items

Hi-DEF subscale	Hi-DEF item examples
Home	3. Switching back and forth between two different activities, such as cooking dinner and answering the phone? 14. Managing your day-to-day finances without making mistakes?
At work	21. Learning new tasks or procedures at work? 24. Responding to changes in your schedule at work?
Driving	28. Driving in an unfamiliar place or unfamiliar route? 29. Staying focused while driving?
Communicating	35. Have difficulty getting your thoughts across in group conversations? 37. Have difficulty managing your emotions in stressful situations?

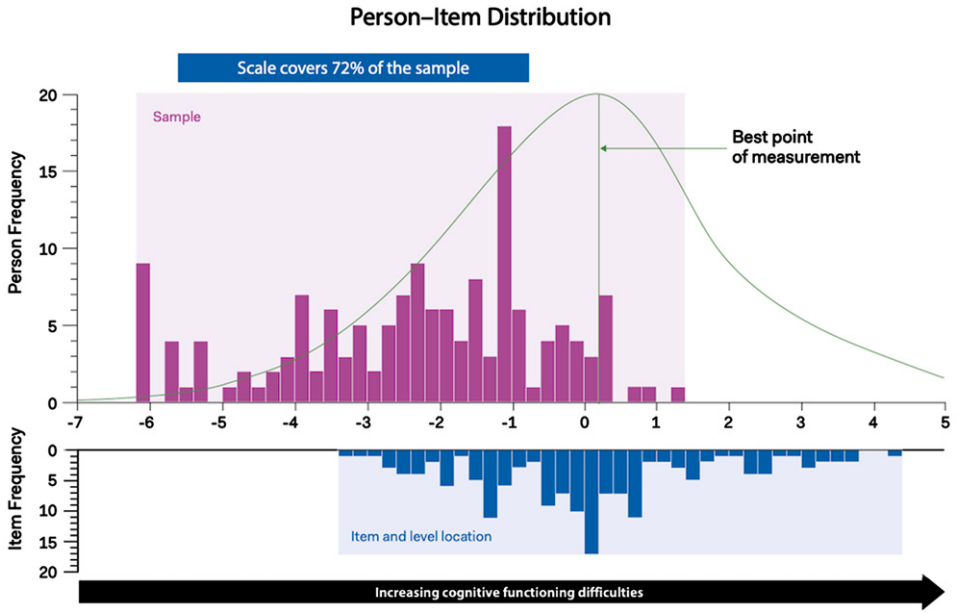

Fig. 1. Hi-DEF scale (total score) targeting plot. The upper histogram represents the sample distribution for the Hi-DEF scale total score whereas the lower histogram represents the scale item threshold distribution plotted on the same linear measurement continuum. This allows a comparison between the range of cognitive functioning difficulties reported in the sample (upper histogram) and the range of cognitive functioning difficulties measured by the items of the Hi-DEF (lower histogram). Overlap between the ranges of the sample and item threshold distributions indicates the instrument is well-matched and able to measure the construct (cognitive functioning difficulties) within the sample accurately. The curve above the upper histogram represents an inverse function of the standard error associated with each person measurement (the peak of the curve indicating the best point of measurement).

was consistent with the draft version across all subscales and the full scale (Table 3). Scale coverage remained sub-optimal for the 'Driving' subscale (49%), adequate for the 'At work' subscale (60%), and good for the 'Home', 'Communicating', and the full scale with coverage ranging from 68% to 72%, respectively (Figs. 1 and 2). As Fig. 1 illustrates, some Hi-DEF items target high levels of difficulties where no patients from the current sample are located; these items are mostly from the 'Driving' and 'Home' subscales as well as two items from the 'At Work' subscale (Fig. 2).

Targeting was also examined with the sample divided by TFC score. On average, the higher the TFC score, the lower the Hi-DEF mean score (indicating fewer impacts of cognitive difficulties on daily functioning), as displayed by the mean logits on Fig. 3. Participants with TFC 9 reported slightly more difficulties on average than those with TFC 8; however, both these groups had a small sample size ($n = 15$ for both). Participants with TFC 10 on average reported more difficulties than TFC 11 or 12, although this was the smallest group of participants ($n = 13$). The mean scores may also have been affected by the participants who reported higher functioning than expected according to their TFC score (located more than one SD below the TFC score group mean): $n = 2$ participants with TFC 10; $n = 3$ participants with TFC

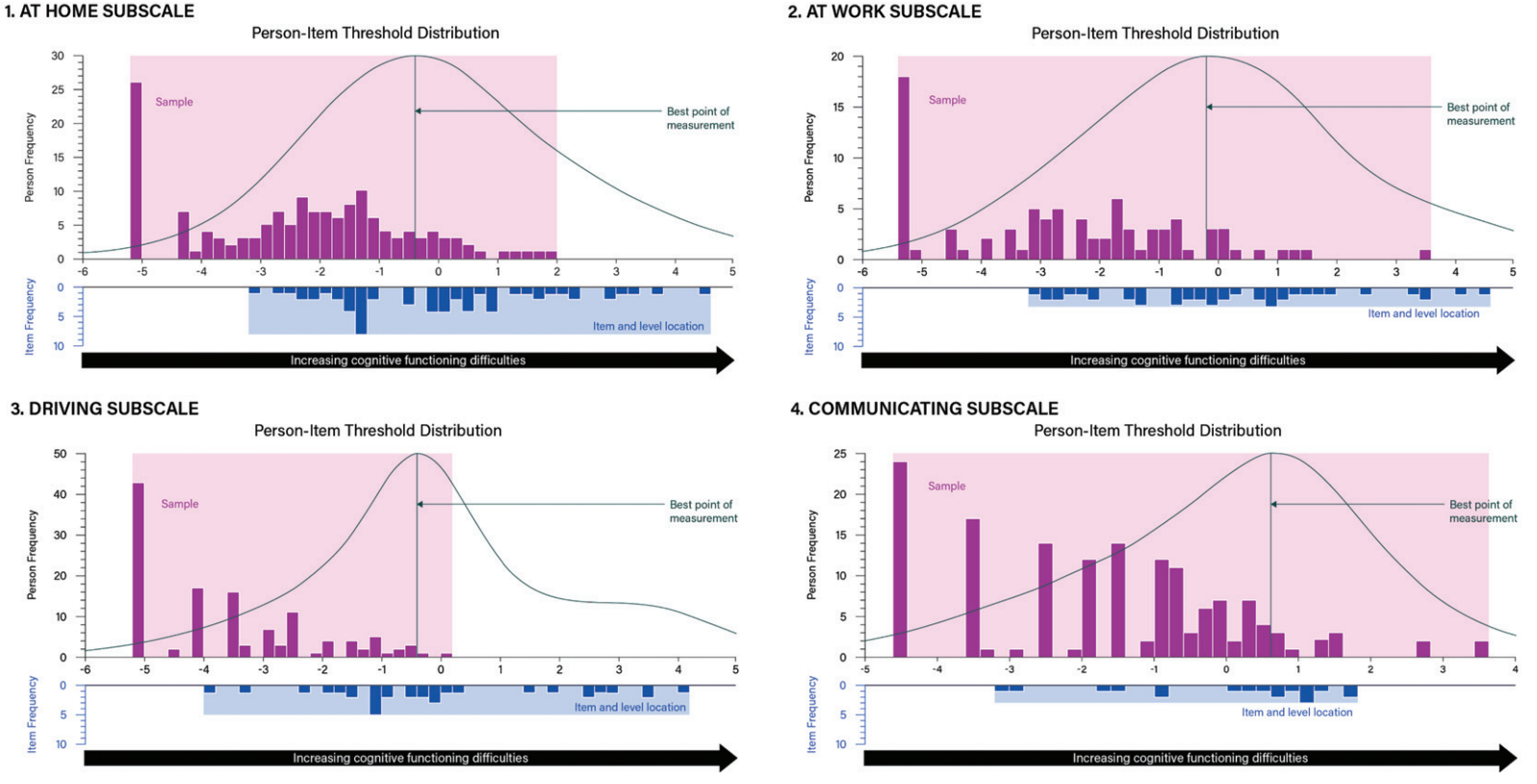

Fig. 2. Targeting plots of Hi-DEF subscales: 1) Home; 2) At work; 3) Driving; 4) Communicating. The upper histogram represents the sample distribution for each Hi-DEF subscale (Home, Work, Driving, and Communicating) whereas the lower histograms represent the item threshold distribution for each subscale plotted on the same linear measurement continuum. This allows a comparison between the range of cognitive functioning difficulties reported in the sample (upper histogram) and the range of cognitive functioning difficulties measured by the items of the Hi-DEF (lower histogram) for each subscale. Overlap between the ranges of the sample and item threshold distributions indicates the instrument is well-matched and able to measure the construct (cognitive functioning difficulties) within the sample accurately. The curve over the upper histogram represents an inverse function of the standard error associated with each person measurement (the peak of the curve indicating the best point of measurement). These figures illustrate the targeting for each of the instrument subscales. The Hi-DEF subscales capture everyday functioning in different environments and may not be applicable to everyone.

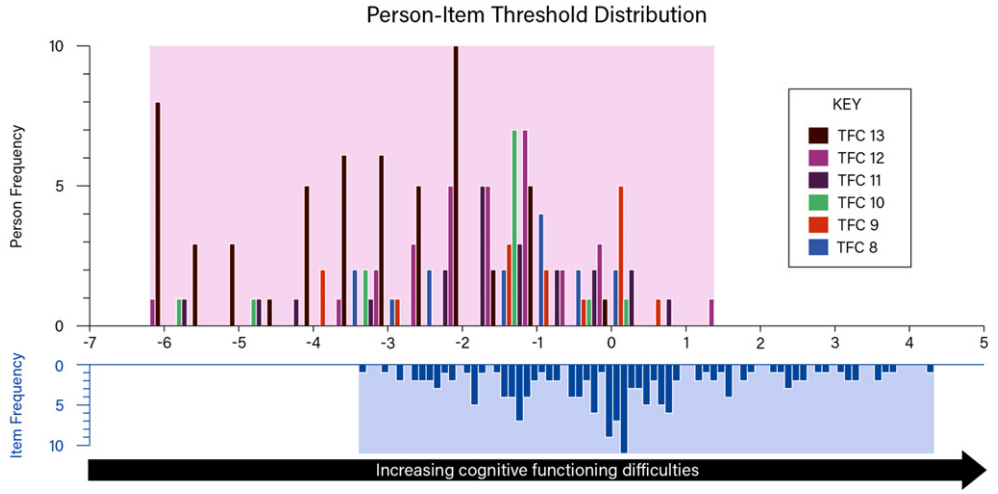

Fig. 3. Hi-DEF full scale targeting plot by TFC score. The upper histogram (variable blocks: colors and patterns indicate TFC score, ranging from TFC 13 to TFC 8, as shown in the key) represents the sample distribution for the Hi-DEF scale, whereas the lower histogram represents the scale item threshold distribution plotted on the same linear measurement continuum.

9; and $n = 3$ with TFC 8. Although some variation within participants with the same TFC score may be expected, it is possible that these participants could have under-reported their difficulties on the Hi-DEF due to a lack of insight into their cognitive impairment and related functioning. All other participants with TFC scores of 10 or lower were located at the lower-functioning end of the continuum (within one SD from the TFC score group mean) which suggests that, other than these eight participants ($n = 8$; 5.3% of the sample), participants' Hi-DEF scores aligned with their TFC scores. This indicates agreement between participants' self-reported cognitive functioning difficulties (Hi-DEF scores) and clinician ratings of their functioning (TFC scores).

Reliability

Reliability assesses the extent to which items can detect differences in cognitive impairment/daily functioning difficulties within a sample, and detect changes over time [32, 33]. Good reliability (the higher the Person Separation Index (PSI) value, the better) indicates a smaller proportion of measurement error in data collected through the Hi-DEF. Reliability of the 40-item Hi-DEF was also consistent with the draft version, with estimated PSI scores ranging between 0.53 ('Driving' subscale) and 0.87 ('Home' subscale), suggesting reliability varied from reasonable to excellent (Table 3). The full Hi-DEF had an estimated PSI of 0.92 (including extreme scores), suggesting excellent reliability.

Item thresholds

Item scoring is justified when patients can discriminate between all response categories (i.e., no difficulty, a little difficulty, a lot of difficulty, etc.) and perceive them as ordered as intended. Along the continuum of cognitive impairment/daily functioning difficulties, there are thresholds where two adjacent response categories are equally likely to be chosen. Ordered thresholds indicate that response options measure distinct categories [34].

The revised 40-item Hi-DEF improved the performance of the item thresholds but did not resolve some of the identified issues, as some item thresholds remained disordered. Specifically, findings indicate the 5-level difficulty response scale worked as intended for 93% (14/15 items) of the 'Home' subscale, 82% (9/11 items) of the 'At work' subscale, and 25% (2/8 items) of the 'Driving' subscale. This suggests participants could not consistently distinguish between the five suggested levels of difficulty for only 1–2 items of the 'Home' and 'At work' subscales, and for 6 items of the 'Driving' subscale. The 'Driving' subscale showed the weakest performance in terms of item thresholds, which was expected since not all of the sample drive. The 4-level frequency response scale of the 'Communicating' subscale displayed no

disordered item response thresholds, indicating participants were able to distinguish between the four implied levels of frequency. In terms of the full Hi-DEF, 75% of the items (30/40 items) displayed ordered item thresholds.

Item fit

Item fit assesses whether Hi-DEF items work together clinically and statistically to form a cohesive continuum of cognitive impairment/daily functioning difficulties. Fit residuals within +/– 2.5 and non-significant chi-square estimates indicate good item fit, which endorses the appropriateness of summing single item responses to obtain a total score [32, 35]. The revised 40-item Hi-DEF improved the cohesiveness of the 'Home,' 'At work,' and 'Communication' subscales while introducing minimal misfit for the 'Driving' subscale, leaving the overall item fit of the full scale unchanged (Table 3). Between 0 and 17% of items in each subscale displayed fit residuals outside the recommended range (+/– 2.5), while review of chi-square only demonstrated marginal misfit in the 'Driving' subscale and the full scale (one item within each scale, corresponding to 12% and 2% of the items in each scale, respectively, which displayed misfit).

Item dependency

Local independence implies that responses to Hi-DEF items are only related due to a shared relationship with the same underlying latent variable, cognitive impairment/daily functioning difficulties, as measured by the Hi-DEF. Any additional relationships between any pair of Hi-DEF items means there is local dependence between those items. Residual correlations > 0.3 are strong evidence of local dependence as this reflects > 9% shared variance, indicating a higher chance that item responses are biased by each other [36, 37]. Item dependency in the 40-item Hi-DEF was consistent for the total and the 'Home' subscale score, with 3% (22/780 item pairs) and 1% (1/105 item pairs) of item pairs, respectively, demonstrating high residual correlations, suggesting minimal dependency. Item pairs of the 'Driving' subscale remained consistently locally independent while item dependency for the 'At work' and 'Communicating' subscales improved with no item pairs demonstrating high residual correlation in the 40-item version (Table 3).

Differential item functioning (DIF)

Item stability (DIF) assesses the extent to which Hi-DEF items are stable and work psychometrically in the same way across subgroups. Non-significant analyses of variance (ANOVAs) indicate no DIF, which endorses that measurements are objective and comparable between subgroups [38]. In the Hi-DEF full scale and subscales, items were stable and are interpreted in the same way across gender and age groups. Sample Hi-DEF items are presented in Table 4.

Person fit

Person fit indicates the extent to which participant response patterns are statistically consistent with the RMT model. Good person fit (up to 5% underfitting) ensures the measurement is valid. The 40-item Hi-DEF demonstrated consistent person fit. The percentage of person misfit for the subscales ranged between 1–5% with marginally higher misfit for the full scale (7%), which improved from the original draft version. These findings are supportive of the validity of sample measurement.

DISCUSSION

The Hi-DEF is a newly-developed PRO instrument designed to assess the impact of cognitive impairment on daily activities in early-stage HD. RMT methods established an optimal set of 40 items, representing tasks of everyday function across four subscales ('Home', 'At work', 'Driving', and 'Communicating'). The Hi-DEF showed good targeting, including targeting by TFC score which was primarily in line with expectation, and excellent reliability, with minimal items showing misfit or dependency, suggesting the scale content is cohesive and unambiguous. These results provide a robust indication of the reliability (as assessed by PSI) and within scale validity (as assessed by item fit, dependency, and person fit) of the Hi-DEF.

In terms of ability of the Hi-DEF to measure impairment across a range (high to low), there were no gaps along the continuum; however, some participants were at the ceiling (i.e., with no difficulties), which is expected since the sample was early stage [TFC mean (SD) = 11.37 (1.72)]. However, since no floor effects were observed and there were many item locations in areas which corresponded to higher level of cognitive functioning difficulties on the con-

tinuum, the Hi-DEF scale may have the potential to pick up deterioration as functioning difficulties increase.

The Hi-DEF helps address a gap in existing instruments for assessing cognitive impairment in HD. It serves as a crucial complement to objective performance-based outcome assessments (PerfOs) [15, 39] and is specifically designed and validated to capture the patients' self-perception and experience of how subtle changes in cognitive function impacts their daily living [15, 39]. Other available HD specific PRO instruments focus on broader HD symptoms, impacts, and health-related quality of life [40–42]. The Hi-DEF is uniquely tailored to measure higher order executive functioning impairment, focusing on different areas/settings of life including home, work, driving, and communication. Thus, the Hi-DEF will provide a more in-depth measurement for cognition, as there are 40 items dedicated to the measurement of cognitive impairment. Finally, the patient voice is of particular importance in rare diseases, such as HD [43]. The Hi-DEF was developed using mixed methods research to ensure that it is patient centered, such that the content is important, relevant, and meaningful to patients and their families [14].

The study findings should be viewed under certain provisos and limitations. Despite rigorous methods and intentional sampling, no information on participants' treatment and medication regimen was available, the sample was predominantly white and non-Hispanic while also highly educated, and premorbid IQ was not assessed. Future research would benefit from more diverse patient samples as well as samples with less education to be more representative of the wider HD community and confirm the Hi-DEF's appropriateness and measurement properties in such populations. Concerning the small sample size, this study provided initial evidence of validation of the Hi-DEF in this population. Further validation will be achieved through the inclusion of the scale in clinical and observational studies in HD. Additionally, since a high proportion of participants were prodromal (TFC score of 13) and presented a limited range of cognitive complaints on both PROs, future studies could explore the use of the Hi-DEF in a more cognitively impaired HD sample. This would allow further investigation of whether Hi-DEF responses are affected by a lack of insight into cognitive issues as cognitive deficits worsen [44]. Moreover, the results of this study should be interpreted with the proviso that apathy was not assessed. Future studies could examine the extent to which apathy influences self-reported Hi-DEF scores. Additional analyses using classical test theory analysis to further explore the reliability and construct validity have been completed, including the ability to discriminate between different levels of functional impairment (based on TFC) and convergent validity with TFC scores, CANTAB scores, and the HD-PRO-TRIAD subscales [45].

In conclusion, the Hi-DEF is a reliable and valid measure which has the potential to facilitate assessment of functional changes associated with cognitive sequelae of HD, with an emphasis on higher-order executive abilities integral to working, driving, and social function. The Hi-DEF is intended for use in both clinical trials and clinical practice and is currently being used in an interventional program which will provide further evidence on the change over time and responsiveness of the Hi-DEF. Psychometric analysis of the Hi-DEF using RMT supports its potential use in clinical research. The use of the Hi-DEF scale in clinical trials may improve measurement of cognitive impairment and its impact on daily functioning, thus supporting the assessments of treatment benefit in HD clinical trials. Future work will involve longitudinal assessment to ascertain its sensitivity to clinical change, and to establish a meaningful change threshold.

ACKNOWLEDGMENTS

We would like to thank all the individuals with HD and their families for their tremendous support, time, and effort participating in this study.

We would like to acknowledge Aaron Koenig (formerly Sage Therapeutics, Inc.) for his contributions to the study. We would like to thank the following investigators who helped recruit for the study: Rajeev Kumar – Rocky Mountain Movement Disorders, Englewood, CO; Jennifer Klapper – Penn Huntington's Disease Center, Philadelphia, PA; Luis Sierra – Beth Israel Deaconess Medical Center, Boston, MA; Jee Bang – Johns Hopkins Medicine, Baltimore, MD; Karen Elta Anderson – Georgetown University, Washington, DC; Deborah Hall – Rush University, Chicago, IL; Susan Perlman – University of California, Los Angeles, CA; Henry Moore – University of Miami, Miami, FL; Danny Bega – Northwestern Medicine, Chicago, IL.

We would also like to thank Aviva Gillman (Modus Outcomes) for help coordinating the study, Flora Mazerolle (Modus Outcomes) for help with data

preparation and Trais Pearson (Modus Outcomes) for scientific writing support.

FUNDING

The study was funded by Sage Therapeutics.

CONFLICT OF INTEREST

Jennifer Petrillo, Ruta Sawant, and Jason Johannesen are employees of Sage Therapeutics, and may have stock/stock options. Sarah Baradaran was an employee of Sage Therapeutics at the time the study was conducted and may have stock. Rebecca Rogers, Sophie Cleanthous, and Stefan Cano are employees of Modus Outcomes, a division of THREAD, which received payment from Sage Therapeutics to conduct this research. Emma Elliott is a former employee of Modus Outcomes (an employee of Modus Outcomes at the time this study was conducted). Logistical support was provided by Boston Strategic Partners, Inc. (funded by Sage Therapeutics).

DATA AVAILABILITY

The data supporting the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy, ethical restrictions, or other concerns. The Hi-DEF is available for licensing and use by the HD and research communities by accessing the ePROVIDE link (https://eprovide.mapi-trust.org/instruments/huntington-s-disease-everyday-functioning-scale).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JHD-240001.

REFERENCES

[1] McColgan P, Tabrizi SJ. Huntington's disease: A clinical review. Eur J Neurol. 2018;25(1):24-34.
[2] Yohrling G, Raimundo K, Crowell V, Lovecky D, Vetter L, Seeberger L. Prevalence of Huntington's disease in the US (954). AAN Enterprises; 2020.
[3] Keum JW, Shin A, Gillis T, Mysore JS, Elneel KA, Lucente D, et al. The HTT CAG-expansion mutation determines age at death but not disease duration in Huntington disease. Am J Hum Genet. 2016;98(2):287-98.
[4] Kwa L, Larson D, Yeh C, Bega D. Influence of age of onset on Huntington's disease phenotype (1885). AAN Enterprises; 2020.
[5] Paulsen JS. Cognitive impairment in Huntington disease: Diagnosis and treatment. Curr Neurol Neurosci Rep. 2011;11(5):474-83.
[6] Paulsen JS, Nopoulos PC, Aylward E, Ross CA, Johnson H, Magnotta VA, et al. Striatal and white matter predictors of estimated diagnosis for Huntington disease. Brain Res Bull. 2010;82(3-4):201-7.
[7] Papoutsi M, Labuschagne I, Tabrizi SJ, Stout JC. The cognitive burden in Huntington's disease: Pathology, phenotype, and mechanisms of compensation. Mov Disord. 2014;29(5):673-83.
[8] Hendel RK, Hellem MNN, Larsen IU, Vinther-Jensen T, Hjermind LE, Nielsen JE, et al. Impairments of social cognition significantly predict the progression of functional decline in Huntington's disease: A 6-year follow-up study. Appl Neuropsychol Adult. 2022. https://doi.org/10.1080/23279095.2022.2073824
[9] Vinther-Jensen T, Larsen IU, Hjermind LE, Budtz-Jørgensen E, Nielsen TT, Nørremølle A, et al. A clinical classification acknowledging neuropsychiatric and cognitive impairment in Huntington's disease. Orphanet J Rare Dis. 2014;9:114.
[10] Diamond A. Executive functions. Ann Rev Psychol. 2013;64(1):135-68.
[11] Jurado MB, Rosselli M. The elusive nature of executive functions: A review of our current understanding. Neuropsychol Rev. 2007;17(3):213-33.
[12] Rabinovici GD, Stephens ML, Possin KL. Executive dysfunction. Continuum (Minneap Minn). 2015;21(3 Behavioral Neurology and Neuropsychiatry):646-59.
[13] Hendel RK, Hellem MNN, Hjermind LE, Nielsen JE, Vogel A. On the association between apathy and deficits of social cognition and executive functions in Huntington's disease. J Int Neuropsychol Soc. 2022;29(4):369-76.
[14] Billet J, Levine A, Johannesen J, Lovell T, Rams A, Gusse E, et al. Patient experiences in early Huntington's disease-qualitative research to inform development of a patient-reported instrument of everyday functioning (P3-11.003). AAN Enterprises; 2022.
[15] Bolink S, Grimm B, Heyligers I. Patient-reported outcome measures versus inertial performance-based outcome measures: A prospective study in patients undergoing primary total knee arthroplasty. Knee. 2015;22(6):618-23.
[16] Carlozzi NE, Victorson D, Sung V, Beaumont JL, Cheng W, Gorin B, et al. HD-PRO-TRIAD™ validation: A patient-reported instrument for the symptom triad of Huntington's disease. Tremor Other Hyperkinet Mov (N Y). 2014;4:223.
[17] Carlozzi NE, Schilling SG, Lai JS, Paulsen JS, Hahn EA, Perlmutter JS, et al. HDQLIFE: Development and assessment of health-related quality of life in Huntington disease (HD). Qual Life Res. 2016;25(10):2441-55.
[18] Fuller RF, P, Lapelle N, Sathe S, Fitzer-Attas C, Guttman M, Goetz C, et al. Functional Rating Scale 2.0 (FuRST2.0), a patient reported outcome measure for Huntington's disease: The importance of the patient voice in scale development [abstract]. Mov Disord 2022;37(suppl 2):Abstract 870.
[19] Petrillo J, Cano SJ, McLeod LD, Coon CD. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. Value Health. 2015;18(1):25-34.

[20] Stover AM, McLeod LD, Langer MM, Chen W-H, Reeve BB. State of the psychometric methods: Patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes. 2019;3:50.
[21] Nguyen TH, Han H-R, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient. 2014;7:23-35.
[22] Billet J, Levine A, Johannesen J, Sawant R, Lovell T, Rams A, et al. Psychometric validation of Huntington's Disease Everyday Functioning (Hi-DEF) Scale –study design and sample characteristics (P3-11.005). Neurology. 2022;98(18_supplement).
[23] Shoulson I, Fahn S. Huntington disease: Clinical care and evaluation. Neurology. 1979;29(1):1-3.
[24] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research (Expanded edition (1980) with foreword and afterword by B.D. Wright, Chicago: The University of Chicago Press, 1980. Reprinted Chicago: MESA Press, 1993. Available from www.rasch.org/books.htm); 1960.
[25] Wright B, Stone M. Best test design: Rasch measurement. Chicago: MESA Press; 1979.
[26] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoeconomics Outcomes Res. 2011;11(5):571-85.
[27] Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiples sclerosis: The role of new psychometric methods. Health Technol Assess. 2009;13(12):1-214.
[28] Petrillo J, Cadavid D, Castrillo-Viguera C, Cleanthous S, Pompilus F, Strzok S, et al. Expanding our understanding of daily life activity impact in patients with multiple sclerosis. 7th Joint ECTRIMS-ACTRIMS Meeting; Paris, France, 2017.
[29] Food and Drug Administration. Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. http://www.fda.gov/downloads/Drugs/Guidances/UCM193282pdf; 2009.
[30] Food and Drug Administration. Qualification Process for Drug Development Tools 2010 [Available from: www.fda.gov/cber/gdlns/prolbl.pdf. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf
[31] Food and Drug Administration. Roadmap to Patient-focused Outcome Measurement in Clinical Trials 2013 [Available from: http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/UCM370174.pdf
[32] Cano SJ, Posner HB, Moline ML, Hurt SW, Swartz J, Hsu T, et al. The ADAS-cog in Alzheimer's disease clinical trials: Psychometric evaluation of the sum and its parts. J Neurol Neurosurg Psychiatry. 2010;81(12):1363-8.
[33] Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, et al. Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods. Alzheimers Dement. 2013;9(1):S4-9.
[34] Andrich D, De Jong J, Sheridan BE. Diagnostic opportunities with the Rasch model for ordered response categories. In Rost J, Langeheine R, editors. Applications of latent trait and latent class models in the social sciences. Waxmann Publishing Co.; 1997. p. 59-70.
[35] Wright BD, Masters GN. Rating scale analysis: MESA press; 1982.
[36] Andrich D. Controversy and the Rasch model: A characteristic of incompatible paradigms? Med Care. 2004;42:I7-I16.
[37] McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? Qual Life Res. 1995;4(4):293-307.
[38] Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess. 1995;7(3):286.
[39] Benedict RH, DeLuca J, Phillips G, LaRocca N, Hudson LD, Rudick R, et al. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. Mult Scler. 2017;23(5):721-33.
[40] Victorson D, Carlozzi NE, Frank S, Beaumont JL, Cheng W, Gorin B, et al. Identifying motor, emotional–behavioral, and cognitive deficits that comprise the triad of HD symptoms from patient, caregiver, and provider perspectives. Tremor Other Hyperkinet Mov (N Y). 2014;4:224.
[41] Hocaoglu M, Gaffan EA, Ho AK. The Huntington's Disease health-related Quality of Life questionnaire (HDQoL): A disease-specific measure of health-related quality of life. Clin Genet. 2012;81(2):117-22.
[42] Clay E, De Nicola A, Dorey J, Squitieri F, Aballéa S, Martino T, et al. Validation of the first quality-of-life measurement for patients with Huntington's disease: The Huntington Quality of Life Instrument. Int Clin Psychopharmacol. 2012;27(4):208-14.
[43] Morel T, Cano SJ. Measuring what matters to rare disease patients–reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. Orphanet J Rare Dis. 2017;12(1):171.
[44] Cacciamani F, Houot M, Gagliardi G, Dubois B, Sikkes S, Sánchez-Benavides G, et al. Awareness of cognitive decline in patients with Alzheimer's disease: A systematic review and meta-analysis. Front Aging Neurosci. 2021;13:697234.
[45] Petrillo J, Elliott E, Rogers R, Cleanthous S, Sawant R, Mazerolle F, et al. Reliability and Construct Validity of the Huntington's Disease (HD) Everyday Functioning (Hi-DEF) Using Classical Test Theory Approach. Abstracts of the 29th Annual Meeting of the Huntington Study Group; November 3–5, 2022; Tampa, Florida.