Journal of Clinical Epidemiology 103 (2018) 51−59

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Classical test theory and item response theory produced differences on estimation of reliable clinical index in World Health Organization Disability Assessment Schedule 2.0

Juan José Mancheño[a], Marcos Cupani[b], Marisa Gutiérrez-López[a], Elena Delgado[c], Enrique Moraleda[c,d], Pilar Cáceres-Pachón[c], Fermín Fernández-Calderón[c,d], Óscar M. Lozano Rojas[c,d,*]

[a]*Community Mental Health Units, Juan Ramón Jiménez Hospital, Huelva, Spain*
[b]*Instituto de Investigaciones Psicológicas, Faculty of Psychology, University of Cordoba, Córdoba, Argentina*
[c]*Department of Clinical and Experimental Psychology, University of Huelva, Huelva, Spain*
[d]*Research Center on Natural Resources, Health and Environment, University of Huelva, Huelva, Spain*

Accepted 5 July 2018; Published online 6 August 2018

Abstract

Objective: World Health Organization Disability Assessment Schedule (WHODAS) 2.0 is currently one of the most used instruments in disability assessment. The objective of this study was to analyze the clinically reliable change of WHODAS 2.0 by applying both Classical Test Theory (CTT) and the Item Response Theory (IRT).

Study Design and Setting: The sample consisted of 179 patients with dual pathology. The standard error of measurement (SEM) was estimated using the CTT and the rating testlet model.

Results: Reliability estimated by Cronbach's alpha provided acceptable values for all domains. The Rasch analysis revealed an adequate capacity to discriminate between people with high and low disability in terms of total scores but not in terms of domains. The SEM varies according to the baseline scores, failing to detect clinically reliable change in patients with lower scores. Kappa coefficients are low for the most of dimensions (except participation) and adequate for total scores.

Conclusion: The use of total WHODAS 2.0 scores may be useful from a clinical perspective; however, more evidence is required for domain scores to support its usefulness. The decision to use the CTT or the IRT impacts in terms of calculating clinically reliable change.

Keywords: Reliable clinical change; WHODAS; Classical test theory; Item response theory; Patient-reported outcome measures; Disability

1. Introduction

The use of Patient Reported Outcome Measures (PROMs) has increased in the clinical and research field during recent years. In general, these instruments assess the impact of treatments on disease as perceived by patients, complementing other indicators that are based on biomarkers. Some studies indicate that the use of PROMs allows for better decision-making in relation to patient interventions [1].

However, the use of such measures in the clinical setting is still limited [2], and there are projects that aim to provide support and guidance for their administration in this context [3,4]. However, one of the barriers to administering these measures concerns the clinical interpretation of the scores

Conflict of interest: The authors declare that they have no conflicts of interest.

Funding: This work was supported by the Fundación Progreso y Salud (call for research projects in Biomedicine and Health Sciences in Andalusia for 2014), project PI-0287-2014.

Ethics: All procedures involved in this study agree with the ethical standards of the institutional and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This study was approved by the ethics committee of the University of Huelva and the hospital center to which the Mental Health Units belong.

Informed consent: Informed consent was obtained from all individual participants included in the study.

* Corresponding author. Dpto. Psicología Clínica, Experimentaly Social, Universidad de Huelva, Campus de 'El Carmen', Avda. Fuerzas Armadas, s/n. 21071, Huelva, España. Tel.: +34 959 21 92 00.

E-mail address: oscar.lozano@dpsi.uhu.es (Ó.M. Lozano Rojas).

https://doi.org/10.1016/j.jclinepi.2018.07.002
0895-4356/

What is new?

- This is the firs study comparing Reliable Change Index obtained by CTT and a Rasch model in the WHODAS 2.0.
- Results show that when CTT or Rasch testlet is used, notable differences have been observed on the Reliable Change Index applied to the WHODAS 2.0 scores.
- Main differences in kappa index of reliable clinical change between CTT and Rasch testlet model are observed in low scores of WHODAS 2.0, where a floor effect is observed.
- The use of the CTT or Rasch testlet model has clinical implications to decide whether patient has improved or worsened.

of PROMs. In this regard, the Consensus-based Standards for the selection of health measurement instruments indicate that providing evidence of the responsiveness and interpretability of the scores can contribute to the applicability of PROMs in clinical practice [5]. Both of these properties are related to an evaluation of the change in scores; however, responsiveness refers to the ability to detect changes in the measured construct and is generally assessed through statistical significance, whereas interpretability refers to the capacity to assign an interpretation to quantitative scores or a change in these scores.

One of the most commonly used statistics for assessing change in patient scores due to the impact of treatment or disease deterioration is the reliable change index (RCI). The RCI evaluates individual change between two defined moments and establishes if the observed differences between the two evaluations can be explained by the measurement error of the instrument or by a real change in the development of the patient [6]. There are different procedures for estimating RCI [6−9], with one of the most widely used being the method proposed by Jacobson and Truax [10].

To calculate the RCI, it is necessary to know the standard error of measurement (SEM), which is generally estimated by applying Classical Test Theory (CTT) [11]. This theoretical approach produces an equal SEM for all the evaluated items and people. That is, the SEM is constant, which implies that subjects with high, medium, and low scores have the same value, although it is acknowledged that the precision of the measures can vary across the continuum underlying the measured construct [12]. This, together with the fact that when applying CTT we obtain an ordinal scale score, has led us to question its usefulness in those contexts where patient change is evaluated as a consequence of the administration of a treatment [12,13]. In contrast to CTT, the Item Response Theory (IRT) brings together a set of psychometric models that, among other properties, provide a measurement error for each person and for each item, as well as a measure of the interval scale [14]. Both of these properties allow for a better interpretation of patients scores observed change [15,16].

From an empirical perspective, relatively few studies have analyzed whether the decision to use either of these psychometric models will have an impact on the RCI. Jabrayilov et al. [17] reported a study using simulated data and concluded that application of CTT or IRT may have advantages and disadvantages depending on the context of use. Moreover, although for tests with at least 20 items the IRT appears to show superior results compared with CTT, there are relatively few discrepancies between the two methods. Brouweret al. [18] also analyzed the RCI of the Beck Depression Inventory-II by applying CTT and IRT to a sample of 104 patients in outpatient treatment. These authors failed to find differences in the classification of the majority of the patients, with the exception of eight subjects that occupied extreme positions on the continuum. This result, therefore, could be taken to reflect the possible impact of ceiling and floor effects on the RCI.

It should be noted that one of the most widely used PROMs in the assessment of disability is the World Health Organization Disability Assessment Schedule (WHODAS) 2.0 [19], which has been adapted to at least 47 languages and administered in 94 countries [20]. This instrument was designed for the assessment of disability from a set of dimensions of the International Classification of Functioning, Disability, and Health (ICF) [19]. WHODAS 2.0 provides information on disability across six domains: cognition (six items), mobility (five items), self-care (four items), getting along (five items), life activities (four items), and participation in society (eight items). Each of these domains can be evaluated independently, although an overall score is also obtained by applying two scoring systems: a simple scoring system, recommended for a clinical setting; or a complex scoring system, based on the application of IRT [21]. From a psychometric perspective, the review by Federici et al. [20] shows that reliability estimated using Cronbach's alpha and the test-retest procedure provides, for the most part, adequate values (with the exception of the self-care domain). Evidence of validity in relation to other variables such as functionality and quality of life has shown the expected theoretical relationships. In contrast, evidence of validity based on the theoretical internal structure of six domains has revealed discrepant results. Furthermore, it should be noted that various authors have reported high ceiling and floor effects [22−24].

From a clinical perspective, WHODAS 2.0 has been widely used in the field of mental health, and the Diagnostic and Statistical Manual of Mental Disorders (DSM)-5 has incorporated the 36-item version as a measure of disability caused by mental disorders [25]. For the interpretation of scores in the clinical context, this instrument has normative scores with information regarding the percentiles [21]. Other studies have provided evidence on

sensitivity to change scores, reporting effect size values from small to moderate depending on the domain assessed [22], whereas other authors have been unable to confirm hypotheses regarding the responsiveness of this scale [24]. In relation to the RCI, Obbarius et al. [26] calculated a value of 9 for the total score of WHODAS 2.0, extracted from the data provided by Chwastieak et al. [27].

Thus, despite the wide use of this instrument—which is expected to increase as a result of its incorporation into DSM-5—no studies have been found that have conducted a specific analysis of the RCI values obtained using WHODAS 2.0. Thus, the present study has the following objectives: (1) to provide RCI scores in a sample of patients with dual pathology (substance use disorders and other mental disorders); and (2) to compare the RCI scores obtained using the estimated SEM from the CTT and IRT models.

Considering previous research, we hypothesize that (1.1) WHODAS 2.0 will be useful for detecting clinically significant change between baseline evaluation and 6 months follow-up in dual patients; (2.1) a high agreement between CTT-estimated scores and IRT-estimated scores is expected for patients with nonextreme scores; and (2.2) for those patients with extreme scores, discrepancies between CTT scores and IRT scores are expected.

2. Method

2.1. Design

Thus study employed a longitudinal observational design, with a baseline assessment followed by another at 6 months after the baseline.

2.2. Participants

The sample was composed of 179 dual pathology patients treated at Huelva Community Mental Health Units (Mental Health public health centers) and Drug Abuse and Dependence centers (the Huelva Province Drug Dependence Service).

There were 148 male participants (82.7%) and 31 female participants (17.3%). The mean age of the participants was 41.23 (SD = 11.1) years for men and 40.48 (SD = 11.9) for women. With respect to educational level, 22.9% had not completed primary education, 45.8% had completed primary education, 27.9% had completed high school or vocational training, and 3.4% had completed university education. At the time of the interview, 45.3% were unemployed and 41.9% were in receipt of benefits for their mental disorders; 8.4% were employed and 3.4% were on sick leave; 69.9% of the patients were single, 8.5% were living with their partner, 19% were separated or divorced, and 2.6% were widowed.

With respect to the drug use profiles of the sample, 45.8% had problems with cocaine use, 40.8% with alcohol, 44.7% with cannabis, and 24.6% with heroin. In relation to comorbid mental disorders, the most prevalent were agoraphobia (48.6%) and major depressive disorder (43.6%). The prevalence of patients with generalized anxiety disorder was 28.5%, as well as patients with psychotic disorders. A lower prevalence was found in patients with antisocial disorder (20.7%), obsessive-compulsive disorder (18.4%), posttraumatic stress disorder (16.2%), social phobia (15.6%), dysthymic disorder (12.8%), and panic disorder (12.3%). With regard to personality disorders, the most prevalent were avoidant disorder (27.9%), borderline disorder (24.6%), and schizotypal disorder (20.7%).

Six months after the baseline evaluation, 96 patients were under treatment and were therefore assessed again. The remaining patients dropped out treatment and it was then not possible to administer the follow-up evaluation. In baseline scores, those patients who continued under treatment showed more negative scores in mobility in comparison to those who dropped out treatment, with statistically significant differences in direct scores ($x_{follow\text{-}up}$ = 9.59, sd = 4.97 vs. $x_{dropout}$ = 8.17, sd = 4.32; t = 2.021; df = 176; P = .045) and logit scores ($x_{follow\text{-}up}$ = −0.64, sd = 0.75 vs $x_{dropout}$ = −0.86, sd = .73; t = 1.989; df = 176; P = .048). For the rest of dimensions, no significant differences were found.

Sample was recruited from January 2016 to August 2017. Follow-up was extended until February 2018.

2.3. Instruments

2.3.1. Sociodemographic/clinical questionnaire

This questionnaire compiled information on the following variables: gender, age, marital status, employment status, and educational level.

2.3.2. Questionnaire for the World Health Organization Disability Assessment (WHODAS 2.0)

The complete version is composed of 36 questions divided into six dimensions: cognition, mobility, self-care, getting along, life activities, and participation [19,21]. WHODAS 2.0 also provides a total disability score corresponding to the sum of all items. Of the two scoring procedures that can be used within WHODAS [19,20], in this study, the simple scoring system was employed. With this procedure, the total score for each dimension corresponds with the sum of scores of each of its items, assigning the value one for the response ‘‘none’’; two for ‘‘mild’’; three for ‘‘moderate’’; four for ‘‘severe’’; and five for ‘‘extreme’’. Because most patients were unemployed, following the instructions of WHODAS 2.0, the total score was 32 items, eliminating those items related to employment. A higher score is indicative of a greater disability.

2.3.3. Mini International Neuropsychiatric Interview

The MINI is a structured diagnostic interview widely used in the field of psychiatry [28]. Through this, the main psychiatric disorders of Axis I can be evaluated. In this

study, we used the Spanish version that evaluates disorders according to the DSM-IV diagnostic criteria, adapted by Ferrando et al. [29].

2.3.4. Personality inventory for DSM-5

The 100-item version was administered [30], which is divided into 25 facets whose combination establishes the personality diagnoses as set out in the DSM-5 25. The items translated into Spanish were extracted from the adaptation of Gutiérrez et al. [31].

2.4. Procedure

The tests were administered by a psychologist with experience in patient assessment. Before starting the period of data collection, the psychologist received specific training for the administration of these tests. The interviews were conducted in the centers where the patients received their treatment, and the patients had previously been informed by the therapist. The patients were also told that the study was unrelated to their therapeutic process. The interviews were conducted in individual sessions in which the psychologist explained the objectives of the research and the participants were informed of the voluntary nature of their participation. They were also told that, unless expressly authorized, the information collected would not be included in their medical history. They were then read the informed consent and asked to sign the form if they wished to participate. Once the patient had signed the consent form, the interview began. This study was approved by the ethics committee of the University of Huelva and the hospital center to which the Mental Health Units belong.

2.5. Data analysis

Using the CTT approach, reliability was analyzed as internal consistency through the Cronbach alpha coefficient, from which the SEM was calculated. The RCI was calculated according to the formula of Jacobson and Truax [10].

$$RC = X_2 - X_1 / S_{diff} = X_2 - X_1 \Big/ \sqrt{2(S_e)^2}$$

where X_2 is the mean of the evaluation scores at 6 months, X_1 is the mean of the baseline assessment scores, and S_e is the standard error of measure.

To calculate RCI with obtained scores by means of IRT, we first applied Rating Scale Model [32], assuming unidimensionality for the total scores, as specified in the guide for users of WHODAS 2.0 [19]. However, a principal component analysis of the residuals conducted with the software Winsteps showed a nonunidimensional structure. Consequently, we carried out a multidimensional Rasch model. Specifically, we conducted a partial-credit testlet model, which is a specific variation of a multidimensional random coefficients multinomial logit model [33]. In this model, it is assumed that each item contributes, on the one hand, to the score in its own factor, and on the other hand to the global score obtained from the items in the test. To estimate the items' parameters, we applied a marginal maximum likelihood (following the Monte Carlo method). After estimating the items' parameters, ability parameters of subject (θ) were obtained using a maximum likelihood procedure. Items fit to model are shown in Table 1. An adequate fit is observed on all items except for items D5.4 and D6.4.

Table 1. Items parameters and fit to the Rasch testlet model

Item	Difficulty level	MNSQ (weighted fit)[a]	CI	T[b]
D1.1	0.318	1.08	0.80, 1.20	0.8
D1.2	0.140	1.08	0.80, 1.20	0.8
D1.3	0.233	1.03	0.80, 1.20	0.3
D1.4	0.409	1.05	0.78, 1.22	0.5
D1.5	0.463	0.95	0.78, 1.22	−0.4
D1.6	0.506	1.15	0.77, 1.23	1.2
D2.1	0.565	1.14	0.76, 1.24	1.2
D2.2	0.819	1.24	0.73, 1.27	1.7
D2.3	1.020	1.36	0.66, 1.34	1.9
D2.4	0.620	1.04	0.74, 1.26	0.3
D2.5	0.514	0.94	0.75, 1.25	−0.5
D3.1	0.826	0.88	0.67, 1.33	−0.7
D3.2	0.941	0.82	0.65, 1.35	−1.0
D3.3	0.754	1.30	0.68, 1.32	1.7
D3.4	0.465	1.22	0.75, 1.25	1.6
D4.1	0.395	0.92	0.80, 1.20	−0.8
D4.2	0.503	0.87	0.78, 1.22	−1.2
D4.3	0.668	1.26	0.72, 1.28	1.7
D4.4	0.369	1.01	0.79, 1.21	0.2
D4.5	0.460	1.06	0.74, 1.26	0.5
D5.1	1.108	1.02	0.71, 1.29	0.2
D5.2	1.174	0.99	0.73, 1.27	−0.0
D5.3	0.806	0.98	0.76, 1.24	−0.1
D5.4	0.820	1.32	0.76, 1.24	2.4
D6.1	0.177	1.00	0.82, 1.18	−0.0
D6.2	0.340	0.96	0.81, 1.19	−0.4
D6.3	0.335	1.11	0.80, 1.20	1.0
D6.4	0.309	1.28	0.80, 1.20	2.5
D6.5	0.116	0.90	0.83, 1.17	−1.2
D6.6	0.369	1.16	0.75, 1.25	1.2
D6.7	0.283	0.91	0.82, 1.18	−1.0
D6.8	0.258	0.91	0.82, 1.18	−1.0

[a] Mean Square (MNSQ) fit values. In weighted fit statistics or infit statistic, the residual is weighted by the information function, which reduces the influence of extreme values.

[b] *t*-Test of the hypotheses ''MNSQ fit statistic is within the interval''. Values greater than two correspond to values outside the interval.

The RCI for the derivations of the IRT was calculated through the following formula [34]:

$$z = \theta_u - \theta_v \Big/ \sqrt{SE_u^2 + SE_v^2}$$

where θ_u and θ_v are the ability parameters of each subject in the first and second assessment, respectively, and $SE_u^2 - SE_v^2$ is the SEM corresponding to each of the two assessment.

In both estimation procedures, RCI values $\geq$1.96 or $\leq$1.96 were considered as clinically reliable change. Finally, the kappa coefficient was applied to establish the level of agreement between both methods.

The CTT analyzes were conducted using the software SPSS version 20 [35] and the Rasch analysis was carried out using Winsteps software version 3.62.1 [36]. Conquest software was used to apply the Rasch testlet model [37].

3. Results

3.1. Reliability and accuracy indicators estimated with CTT and Rasch testlet model

Table 2 shows the different indicators of WHODAS 2.0 scores—reliability and accuracy—estimated with CTT and Rasch testlet model. It is observed that the floor effect exceeds 15%, recommended as acceptable [38] in the domains of "mobility," "self-care," "getting along," and "life activities". The internal consistency estimated through the Cronbach alpha coefficient shows acceptable values for all domains, with the lowest value corresponding to the self-care domain (α = .71). The highest SEM value is observed for the total scale because this measure is the one with the greatest dispersion (SD = 25.96).

The analysis using the Rasch testlet model reveals that the mean SEM values as a function of the percentile in the score corresponding to the baseline. The highest values are found in the patients with the lowest score ($<P_{25}$), coinciding with domains in which there is a greater floor effect. The last three columns show the mean differences in scores for people who have improved. It can be seen that for all domains except for mobility and self-care, changes are observed for people with scores higher than P_{75}. People with clinically reliable changes with values lower to P_{75} were only observed for the total scores of the WHODAS 2.0.

Fig. 1 shows the relationship between estimated ability and SEM. As it can be seen in this figure, SEM is substantially lower for scores close to 0 logits.

3.2. Comparison of the scores between baseline and follow-up assessment

Table 3 displays the comparison between the baseline and follow-up scores for each of the dimensions of WHODAS 2.0. The comparisons were made for the raw scores obtained by the simple scoring system according to WHODAS 2.0 manual [19], as well as for the ability scores (θ) of a logits scale, after applying the Rasch testlet model. The results indicate that there are no statistically significant differences in any of the domains or in the total score, regardless of whether the scores are estimated with the CTT or the Rasch testlet model. Moderate values for the correlation coefficients between baseline and follow-up were found.

3.3. RCI according to CTT and Rasch testlet model, and agreement on the classification of patients

Table 4 shows the percentages of patients presenting clinically reliable change. When applying CTT, the percentage of people with a change ranges from 13.5% in the "self-care" domain to 39.2% in the "life activities" domain. The clinically reliable change in the total score calculated using the CTT is 47.5%. Applying the Rasch

Table 2. Reliability and accuracy indicators estimated with CTT and rating scale model for WHODAS 2.0 domains and total score

	CTT							Rasch testlet model						
									Mean of standard SEM:			Mean scores of patients that improve with reliable change		
	Floor effect	Ceiling effect	Max.–Min.	P_{25}–P_{75}	Alpha	SEM	Reliable change for people who improve	Max.–Min.	$<P_{25}$	P_{25}–P_{75}	$P_{75}>$	$<P_{25}$	P_{25}–P_{75}	$P_{75}>$
Cognition	12.8	0.6	6–30	9–19	.84	2.62	>7.2	−2.14 to 1.65	0.77	0.31	0.35	-	-	0.50
Mobility	36.3	0.8	5–25	5–11	.82	2.06	>5.69	−1.54 to 1.93	0.93	0.43	0.35	-	-	-
Self-care	49.2	-	4–28	4–7	.71	1.72	>4.75	−0.93 to 1.28	0.88	0.49	0.37	-	-	-
Getting along	30.2	2.2	5–25	5–13	.77	2.41	>6.66	−1.50 to 1.23	0.95	0.39	0.33	-	-	0.68
Life activities	31.8	-	4–20	4–12	.93	1.45	>4.01	−1.41 to 1.56	0.89	0.43	0.42	-	-	0.36
Participation	5.6	-	8–39	12–24	.80	3.78	>10.45	−2.38 to 1.19	0.59	0.26	0.27	-	-	−0.08
Total score	1.7	-	32–147	44.25–82	.94	6.82	>18.85	−3.66 to 1.10	0.40	0.15	0.13	−1.53	−0.53	−0.01

Abbreviations: CTT, classical test theory; SEM, standard error of measurement.

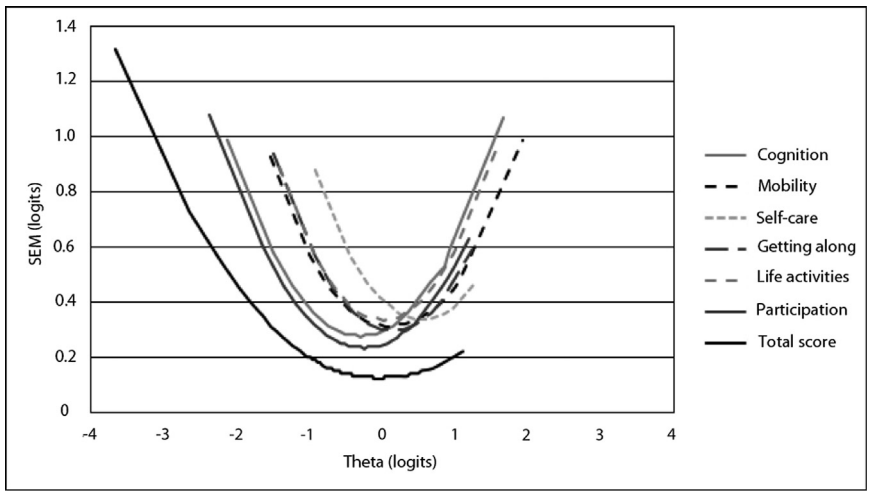

Fig. 1. Relationships between ability persons and SEM. SEM, standard error of measurement.

testlet model shows that the percentage of patients presenting clinically reliable change in the domains ranges from 2.1% (mobility and self-care) to 13.4% (cognition). In the total score, 42.3% of patients with clinically reliable change were observed.

Comparing the percentages of both psychometric approaches, it is clear that for all domains, the CTT produces higher percentages of patients with reliable change in comparison with the IRT. Remarkable differences are observed in the domain ''life activities,'' in which the reliable change between psychometric theories is 34%. For WHODAS 2.0 total scores, the percentage of patients with clinically reliable change is higher than 40% for both psychometric theories. However, when applying CTT an increase of 5% in clinically reliable change is observed in comparison to IRT.

Furthermore, following the Landis and Koch classification for the kappa coefficient [39], the values for all scores are almost perfect for the domains of ''participation'' and ''total scores''. In contrast, a poor or slight agreement for the rest of dimensions is detected. Furthermore, kappa values vary according to the analyzed percentiles. Specifically, the higher agreement in most dimensions is observed between percentiles 25–75. In the case of the domains ''life activities'' and ''participation,'' the higher agreement is observed for percentile 25 or higher. Moreover, for the

Table 3. Comparison between baseline and follow-up on raw scores and θ (logits)

	Baseline		Follow-up						
	M	SD	M	SD	$r_{pre-post}$	t	df	*P*	d
Raw scores									
Cognition	14.14	6.21	14.81	6.36	.487	0.90	95	.372	0.11
Mobility	9.59	4.97	9.91	5.53	.607	0.65	95	.515	0.06
Self-care	6.55	3.59	6.84	3.89	.594	0.85	95	.400	0.08
Getting along	10.02	5.09	10.31	5.58	.582	0.58	95	.561	0.05
Life activities	8.50	4.51	9.21	5.33	.510	1.43	95	.156	0.14
Participation	18.76	7.48	19.90	8.43	.609	1.57	95	.120	0.14
Total score	67.61	25.22	70.98	29.14	.651	1.44	95	.153	0.12
θ (logits)									
Cognition	−0.66	0.80	−0.54	0.71	.411	1.38	95	.172	0.16
Mobility	−0.64	0.75	−0.63	0.83	.586	0.107	95	.915	0.01
Self-care	−0.36	0.64	−0.33	0.67	.487	0.466	95	.643	0.05
Getting along	−0.56	0.74	−0.52	0.77	.560	0.565	95	.574	0.05
Life activities	−0.52	0.71	−0.41	0.89	.458	1.233	95	.220	0.14
Participation	−0.69	0.71	−0.60	0.74	.561	1.273	95	.206	0.12
Total score	−0.70	0.81	−0.63	0.86	.627	0.854	95	.395	0.08

Abbreviations: M, mean; SD, standard deviation.

Table 4. Percentage of reliable clinical change in scores estimated according to CTT and Rasch testlet model, and kappa congruence index

	CTT			Rasch testlet model						
	Improve	Worsen	Change	Improve	Worsen	Change	Kappa total	Kappa P 25 scores	Kappa P 25–75 scores	Kappa $P > 75$ scores
Cognition	17.5	18.6	36.1	7.3	5.2	12.5	.10	0	.06	.23
Mobility	11.3	12.4	23.7	0	2.1	2.1	.14	0	.39	0
Self-care	7.2	6.3	13.5	0	2.1	2.1	.21	.38	.64	0
Getting along	6.2	11.3	17.5	2.1	1	3.1	.29	0	.25	.39
Life activities	15.5	23.7	39.2	3.1	2.1	5.2	.18	.29	.09	.18
Participation	4.1	12.4	16.5	3.1	6.3	9.4	.69	.83	.60	.72
Total scores	20.6	26.8	47.4	19.8	21.9	41.7	.81	.59	.85	.80

Abbreviation: CTT, classical test theory.

domain ‘‘getting along’’ the higher kappa value is observed for those scores higher than percentile 75.

4. Discussion

The general aim of the present study was to provide new psychometric evidence that could be helpful in the clinical interpretation of the WHODAS 2.0 scores in patients with mental disorders. Although many studies have so far analyzed psychometric properties in terms of evidence of reliability and validity, this is the first study to provide a detailed analysis of the RCI with this scale. In addition, given that this scale has shown evidence of ceiling and floor effects [40,41], we decided to analyze the RCI by estimating the SEM using both the CTT and Rasch testlet model.

The reliability analysis with CTT provided adequate Cronbach alpha values that are consistent with those found in previous studies [20]. IRT models have been less widely applied to this scale. When applying Rasch testlet model, we found that two of 32 items (item D5.4: ‘‘Getting your household work done as quickly as needed?’’ and item D6.4: ‘‘How much time did you spend in your health conditions or its consequences?’’) did not show adequate fit values. These two items, which represent 6.25% of test items, exceed the acceptable criterion of 5% misfit items allowed for a test [42]. These findings are not consistent with those by Üstün et al. [19], who found adequate fit values for the all items of the WHODAS 2.0 applying the partial credit model. Nevertheless, they pointed out that the item ‘‘getting your household work done as quickly as needed?’’ had to be recoded. Moreover, Galindo-Garré et al. [43] applied the Partial Credit model in a sample of patients with schizophrenia spectrum disorders. They found that some items (different from items observed in our study) were outside the permissible range of infit and outfit statistics.

The initial examination of items D5.4 and D6.4 in our sample did not let us to establish a hypothesis about the lack of fit that could be further investigated. In light of previous results, along with our findings, we consider that future results applying IRT models to the WHODAS 2.0 should be interpreted with caution. Furthermore, it should be considered that estimating scores through IRT is a more complex process than doing it by means of CTT. Also, interpreting raw scores is more intuitive than interpreting theta values (in logits scale). Consequently, despite from a psychometric perspective using IRT it might be an advantage [14], the lack of fit when IRT models are applied and the complexity in obtaining and interpreting the scores becomes a difficulty to its use in clinical contexts.

Second, and as a highlight of the application of the Rasch testlet model to this scale, it has been shown how the SEM varies according to patient scores. From a psychometric perspective, the SEM of the scores below the 25th percentile is higher than the scores above the 75th percentile. Specifically, it is observed that their values are twice higher in most domains. Considering total scores, the SEM of the scores for 25th percentile is three times higher to the SEM of the scores above 75th percentile. In addition, it is observed that the SEM of the total scores is significantly lower than the SEM observed in each of the domains. However, these values must be interpreted within the plausible scores range. For example, the SEM obtained for scores located below 25th percentile of total score is higher than the SEM found in the domains of ‘‘mobility’’ and ‘‘self-care’’ for scores positioned above the 75th percentile. Nevertheless, in the first case, there are patients with clinically reliable changes; this is not observed in those patients with scores above the 75th percentile of two dimensions mentioned. This can be explained because, although the SEM is higher, the score range has even wider values below the 25th percentile of total score (range for 25th percentile total scores = 2.21 logits; range for mobility scores higher than 75th percentile = 1.42 logits; range for self-care scores higher than 75th percentile = 0.30 logits).

From a clinical perspective, these SEM values have an impact in two ways. First, it has been shown that WHODAS 2.0 scores do not allow for the detection of clinically reliable changes in the domains for patients with scores that fall below the 75th percentile. Thus, for patients with these

scores, it is advisable to use other disabilities instruments that have shown adequate sensitivity. For those patients who had baseline scores below 75th percentile, clinically reliable improvements are detected only in total scores of WHODAS 2.0. Consequently, the clinically reliable change observed after applying IRT model to the domains affects to a reduced proportion of patients (the highest proportion is observed in the domain cognition, 12.5%). In contrast, when applying CTT model, the proportion of clinically reliable change affects to a higher percentage of patients. Thus, regarding our fist hypothesis, it must be considered that the capacity of WHODAS 2.0 to detect clinically reliable changes varies according to the use of CTT or IRT psychometric models.

Second, when comparing the percentage of patients with clinically reliable change when applying CTT and Rasch testlet model, differences in all domains except for ''participation'' and ''total scores'' are detected. It is also observed that for those domains in which a high floor effect exists, the kappa value indicates low agreement between both psychometric theories. In addition, the higher agreement for most domains is observed for those patients whose scores range between percentiles 25−75. This may be related to the fact that in this score range the estimated SEM values with Rasch testlet model are the lowest. Therefore, a higher proportion of patients is classified as reliable change (improve or worsen). This high variability, which is to some extent consistent with the observed changes with CTT, is causing increments in kappa values. This result partially differs with previous research comparing reliable clinical change with IRT and CTT [19,20], and it is not consistent with the hypothesis in our study. Specifically, in contrast to previous research [19,20], we did not find high agreement between the estimated scores when applying CTT and IRT. However, our results are consistent with those by Brouwer et al. [18], who also found discrepancies between both psychometric theories for patients with extreme scores.

On the other hand, various authors have analyzed responsiveness to change through statistical significance and effect sizes [21,22,27]. In this study, no statistically significant differences were found between baseline and follow-up. Therefore, the conclusion would be that no effect of treatment was found in the WHODAS 2.0 scores. However, an analysis based on clinically reliable change for total scores would result in different conclusions. In particular, the results show that when applying the CTT, the percentage of subjects who show a change in their status is equal to 47.4% (when applying the IRT, this is at least 41.7%). We consider this result to be of particular relevance, given that WHODAS 2.0 is one of the most widely used scales in the assessment of disability, and conclusions that are drawn only from statistical significance could lead to misinterpretation. These results also support the view of Coon and Cook [44], highlighting the importance that research community provides evidence of clinical significance when using PROMs, rather than reporting statistical significance only.

Although the authors of the present study consider that the results are of interest for the application of WHODAS 2.0 in particular, and for the clinical use of PROMs in general, it is necessary to bear in mind some limitations. The main limitation concerns the loss of patients during the follow-up phase. It must be noted that the follow-up evaluation was carried out only with the 96 patients who were under treatment after 6 months from baseline evaluation. Considering this sample size, it has been observed that floor effect affects to the reliability of clinical change when CTT or IRT are used. Thus, for a suitable kappa estimation, a higher sample size would be necessary. In this regard, our results must be considered as a first approach, and further research with large sample size is needed to better achieve sound conclusions. Moreover, as WHODAS 2.0 is an instrument for evaluating dysfunction, it was not designed to detect changes in people with a ''normal'' functioning. For this reason, it should be considered that this scale does not necessarily present an incorrect functioning.

A further limitation is related to the procedure used to calculate the scores. As already indicated, WHODAS 2.0 has two scoring systems. In this study, we employed only the simple scoring system. The complex scoring system, on the other hand, carries a weight of the scores, so the results seen in the present study are not generalizable to instances in which the complex scoring system is used. Thus, further research is necessary to complement the present findings.

References

[1] Bingham CO, Noonan V, Auger C, Feldman DE, Ahmed S, Barlett SJ. Montreal Accord on patient-reported outcomes use series-paper 4: patient-reported outcomes can inform clinical decision making in chronic care. J Clin Epidemiol 2017;89:136−41.
[2] Hatfield D, McCulough L, Frantz SH, Krieger K. Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. Clin Psychol Psychotherapy 2010;17:25−32.
[3] Department of Health. Our NHS, our future: NHS next stage review. Interim report 2007. Available at http://www.nhshistory.net/darzi-interim.pdf. Accessed February 3, 2018.
[4] Department of Health. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs) 2009. Available at http://www.healthcare-today.co.uk/doclibrary/documents/pdf/152_guidance_on_routine_collection.pdf. Accessed February 3, 2018.
[5] Mokkink L, Prinsen C, Bouter L, de Vet H, Terwee C. The Consensus-based Standards for the selection of health measurement instruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther 2016;19:105−13.
[6] Wise EA. Methods for analying psychotherapy outcomes: review of clinical significance, reliable change, and recommendations for future directions. J Pers Assess 2004;82:50−9.
[7] Speer DC. Clinically significant change: Jacobson and Truax (1991) revisited. J Consult Clin Psychol 1992;60:402−8.
[8] Speer DC, Greenbaum PE. Five methods for computing significant individual client change and improvement rates: support for an individual growth curve approach. J Consult Clin Psychol 1995;63:1044−8.

[9] Hsu LM. Reliable changes in psychotherapy: taking into account regression toward the mean. Behav Assess 1989;11:459−67.
[10] Jacobson NS, Truax P. Clinical significance: a statistical ap- proach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12−9.
[11] Montero M, Iraurgi I, Matellanes B, Montero JM. Use of the reliable change index to evaluate the effectiveness of clinical interventions: application of an asthma training program. Aten Primaria 2015;47: 644−52.
[12] Petrillo J, Cano S, McLeod L, Coon C. 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;18:25−34.
[13] Grimby G, Tennat A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med 2012;44:97−8.
[14] Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
[15] Reise S, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol 2009;5:27−48.
[16] Chang C, Reeve B. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005;28: 264−82.
[17] Jabrayilov R, Emons WH, Sijtsma K. Comparison of classical test theory and item response theory in individual change assessment. Appl Psychol Meas 2016;40:559−72.
[18] Brouwer D, Meijer R, Zevalkink J. Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. Psychotherapy Res 2013;23:489−501.
[19] Üstün TB. Measuring health and disability: manual for WHO disability assessment schedule WHODAS 2.0. Geneve: World Health Organization; 2010.
[20] Federici S, Bracalenti M, Meloni F, Luciano J. World Health Organization disability assessment Schedule 2.0: an international systematic review. Disabil Rehabil 2016;39:2347−80.
[21] Üstün T, Chatterji S, Kostanjsek N, Jürgen R, Kennedy C, Epping-Jordan J, et al. Developing the world health organization disability assessment Schedule 2.0. Bull World Health Organ 2010;88: 815−23.
[22] Garin O, Ayuso-Mateos JL, Almansa J, Nieto M, Chatterji S, Vilagut G, et al. Validation of the world health organization disability assessment schaedule, WHODAS-2 in patients with chronic diseases. Health Qual Life Outcomes 2010;19:51.
[23] Guilera G, Gómez-Benito J, Pino O, Rojo E, Vieta E, Cuesta MJ, et al. Disability in bipolar I disorder: the 36-item world health organization disability assessment Schedule 2.0. J Affect Disord 2015; 174:353−60.
[24] Moen VP, Drageset J, Eide G, Klollerud M, Gjesdal S. Validation of world health organization assessment Schedule 2.0 in specialized somatic rehabilitation services in Norway. Qual Life Res 2017;26: 505−14.
[25] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Madrid: Panamericana; 2013.
[26] Obbarius A, Massakkers L, Baer L, Clark DM, Crocker A, Beurs E, et al. Standadization of health outcomes assessment for depression and anxiety: recommendations from the ICHOM Depression and Anxiety Working Group. Qual Life Res 2017;26:3211−25.
[27] Chwastiak LA, von Korff M. Disability in depression and back pain: evaluation of the World Health Organization Disability Assessment Schedule (WHODAS II) in a primary care setting. J Clin Epidemiol 2003;56:507−14.
[28] Sheehan D, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropschiatric Interview (M.I.N.I): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59:22−33.
[29] Ferrando L, Bobes J, Gibert J, Soto M, Soto O. MINI entrevista neuropsiquiátrica internacional. Madrid: Instituto IAP; 2000.
[30] Maples J, Carter N, Few L, Crego C, Gore W, Samuel DB, et al. Testing whether the DSM-5 personality disorder trait model can be measured with a reduced set of ítems: an ítem response theory investigation of the Personality Inventory for DSM-5. Psychol Assess 2015;27:1195−210.
[31] Gutiérrez F, Aluja A, Peri JM, Calvo N, Ferrer M, Gutiérrez-Zotes JA, et al. Psychometric properties of the Spanish PID-5 in a clinical and a community sample. Assessment 2015;24:326−36.
[32] Wright BD, Masters GN. Rating scale analysis. Chical, IL: MESA Press; 1982.
[33] Adams RJ, Wilson MR, Wang WC. The multidimensional random coefficients multinomial logit model. Appl Psychol Meas 1997;21:1−23.
[34] Guo J, Drasgow F. Identifying cheating on unproctored Internet tests: the z-test and the likelihood ratio test. Int J Selection Assess 2010;18: 351−64.
[35] IBM Corp. Released. IBM SPSS statistics for windows, Version 20.0. Armonk, NY: IBM Corp; 2011.
[36] Linacre JM. A user's guide to winsteps ministep: Rasch-model computer programs. Chicago: Winsteps; 2007.
[37] Wu ML, Adams RJ, Wilson MR. Conquest: generalized item response modeling software [Computer software and Manual]. Camberwell, Austraila: Australia Council for Educational Research; 1998.
[38] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34−42.
[39] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[40] Magistrale G, Pisani V, Argento O, Chiara CI, Bozzali M, Cadavid D, et al. Validation of the world health organizarion disability Asssessment Schedule II (WHODAS II) in patients with multiple sclerosis. Mult Scler 2015;21:448−56.
[41] De Wolf A, Tate R, Lannin N, Middleton J, Lane-Brown A, Cameron I. The World Health Organization Disability Assessment Scale, WHODAS II: reliability and validity in the measurement of activity and participation in a spinal cord injury population. J Rehabil Med 2012;44:747−55.
[42] Bond TG, Fox CM. Applying the rasch model: fundamental measurement in the human sciences. New Jersey: Lawrence Erlbaum; 2001.
[43] Galindo-Garré F, Hidalgo MD, Guiler G, Pino O, Rojo JE, Gómez-Benito J. Modeling the world health organization disability assessment Schedule II using non-parametric item response models. Int J Methods Psychiatr Res 2015;24:1−10.
[44] Coon C, Cook K. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 2018;27:33−40.