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 xxx

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