Author(s): Saltychev, Mikhail; Katajapuu, Niina; Bärlund, Esa; Laimi, Katri

Title: Psychometric properties of 12-item self-administered World Health Organization disability assessment schedule 2.0 (WHODAS 2.0) among general population and people with non-acute physical causes of disability : systematic review

Year: 2021

Version: Accepted version (Final draft)

Please cite the original version:

Saltychev, M., Katajapuu, N., Bärlund, E., & Laimi, K. (2021). Psychometric properties of 12-item self-administered World Health Organization disability assessment schedule 2.0 (WHODAS 2.0) among general population and people with non-acute physical causes of disability : systematic review. Disability and Rehabilitation, 43(6), 789-794. https://doi.org/10.1080/09638288.2019.1643416

PSYCHOMETRIC PROPERTIES OF 12-ITEM SELF-ADMINISTERED WORLD HEALTH ORGANIZATION DISABILITY ASSESSMENT SCHEDULE 2.0 (WHODAS 2.0) AMONGST GENERAL POPULATION AND PEOPLE WITH NON-ACUTE PHYSICAL CAUSES OF DISABILITY – SYSTEMATIC REVIEW

Short title: Psychometrics of self-administered 12-item WHODAS 2.0

ABSTRACT

Objective

WHODAS 2.0 is a unified scale to measuring disability across diseases, countries, and cultures. The objective was to explore the available evidence on the psychometric properties of 12-item self-administered WHODAS 2.0 amongst a general population and people with non-acute physical causes of disability.

Methods

Five databases Medline, Embase, Web of Science, Scopus and PsycINFO were searched for papers related to the validity, reliability, responsiveness, minimal clinically important difference or minimal detectable change of 12-item self-administered WHODAS 2.0. In order to avoid missing any potentially relevant studies, the search clauses were left as generic as possible and the refining search was conducted manually. As the review was focusing on chronic physical disorders and general adult population, major psychiatric diagnoses, acute traumas, other acute conditions (e.g. postpartum or pregnancy), hearing loss, progressive neurological disorders, and age <19 years were excluded. The relevancy of the studies was assessed by two independent reviewers.

Results

The 14 out of 191 observational studies were considered relevant. The sample sizes varied from 80 up to 31,251 participants. Great diversity was observed in the participants' health problems. The Cronbach's alpha was high – up to 0.96. The correlations between WHODAS 2.0 and other disability scales were high. Substantial floor without ceiling effect was reported by two studies. Exploratory factor analysis resulted in a multidimensional structure – up to five factors. The discriminative ability and test-retest reliability of the scale was good.

Conclusions

It seems, that the 12-item self-administered WHODAS 2.0 is internally consistent and a reliable scale demonstrating overall good correlation with other measures of disability. However, it appears that it is a multidimensional scale and its total score may represent different combinations of several contributing factors. Thus, the 12-item WHODAS 2.0 can be more reliable when creating a person's functional profile formed by the 12 individual item scores instead of a single total sum.

KEYWORDS

disability evaluation; international classification of functioning, disability and health; functioning; psychometrics; reproducibility of results; consistency; floor effect; ceiling effect; whodas

INTRODUCTION

The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) is an ambitious attempt made by WHO to introduce a unified scale to measuring disability across diseases, countries, and cultures [1, 2]. WHODAS is based on the International Classification of Functioning, Disability and Health (ICF), producing standardized numeric disability levels and profiles. The comprehensive 36-item WHODAS 2.0 has been developed in order to describe six latent constructs: cognition, mobility, self-care, getting along, life activities, and participation. In theory, these six 'sub'-constructs (or 'domains' in the terms of ICF) should be able to explain the broader concept of 'general disability'. The number of indicators – WHODAS 2.0 items – varies from four to eight for each of the six latent constructs. The 12-item WHODAS 2.0 has been derived from the 36-item version to provide a briefer tool for assessing overall functioning in surveys or health-outcome studies. Two items for each of the six latent factors have been included in the 12-item WHODAS 2.0. The 12-item version has been found to be reliable, and has been reported to explain 81% of the overall variance of results of the 36-item WHODAS [1]. The total score of WHODAS 2.0 is scored either by using an item response theory or as the simple sum of scores assigned to each of the items.

The WHODAS 2.0 is available in several versions: 36-, 24+12- and 12-item questionnaires in self-, interviewer- and proxy-administered forms. While the full 36-item version has more commonly been used, the shorter 12-item WHODAS 2.0 has raised a great interest among clinicians and researchers as an easy-to-use short indicator of disability, sometimes called a WHODAS 'screener'. About one third of all papers identified by a recent review on WHODAS 2.0 has employed a 12-item version [3].

The psychometric properties of 36-item WHODAS 2.0 has extensively been studied [2]. Overall, it has been described as a consistent, reliable, and unidimensional tool. Instead, the knowledge on the 12-item version's psychometrics is scarce. Previous research has often assumed that psychometric properties of the 12-item version are fully inherited from its more comprehensive 36-item form. For example, several studies have conducted confirmatory factor analysis of the 12-item WHODAS 2.0 based on a presumption of unidimensionality and hierarchical structure (one common factor 'disability' and six subfactors regarding different dimensions of disability) demonstrated by a 36-item [4, 5]. However, one could expect that excluding 24 out of 36 items might affect psychometrics substantially. In other words, it is uncertain how well a 12-item version is able to reproduce the psychometric properties of 36-item WHODAS 2.0.

The research on psychometrics of the 12-item WHODAS 2.0 is scattered across its different forms and diverse populations of interest Firstly, the psychometrics of all three 12-item forms – self-, proxy and interviewer-administered – have sometimes been reported as the properties of a general '12-item WHODAS', even though the psychometrics of self-reported form might differ from proxy- or interviewer-administered assessments.

Secondly, the research on the subject is scattered across numerous relatively small samples with different settings and diagnostic profiles. It is true, that WHODAS 2.0 is a tool that should work, in theory, in any diseases and

settings, but this assumption should first be confirmed by comparing the psychometric properties of the WHODAS 2.0 between large samples involving similar conditions and analogous settings.

The objective of this study was to explore the available evidence on the psychometric properties of the 12-item self-administered WHODAS 2.0 amongst a general population and people with non-acute physical causes of disability.

METHODS

Inclusion and exclusion criteria

Inclusion: Papers (including short communications and letters to editor, excluding conference proceedings, theses etc.) published in academic peer-reviewed journals. No restrictions on time of publication or language.

Exclusion: Major psychiatric diagnoses, acute traumas, other acute conditions (e.g. postpartum or pregnancy), hearing loss, progressive neurological disorders, age <19 years.

Databases: Medline, Embase, Web of Science, Scopus, PsycINFO.

Outcome: Psychometric properties of WHODAS 2.0 understood as any property of WHODAS 2.0 related to its validity, reliability, responsiveness, minimal clinically important difference or minimal detectable change, or respective.

Data sources and searches

The MEDLINE (via PubMed), Embase, Web of Science, Scopus, and PsycINFO databases were searched in January 2019. The search clauses are presented in Table 1. In order to avoid missing any potentially relevant studies, the search clauses were left as generic as possible and the refining search was conducted manually. The references of identified articles and reviews were also checked for relevancy.

Study selection

Two independent reviewer teams (NK + EB vs. MS) screened titles and abstracts of articles and assessed the full texts of potentially relevant studies (Figure 1). Disagreements between the reviewers were resolved by consensus or by a third reviewer (KL). The methodological quality of the included trials was not rated.

Data extraction

The potentially relevant data were extracted from the records by one reviewer using a predefined structured form including title, first author, year of publication, country of origin, study settings, participants' main diagnoses if specified, sample size, gender distribution, participants' age, main psychometric measures used, main quantitative results, and the conclusions drawn by the original authors.

RESULTS

Search results

The search resulted in 191 records. Of them, 148 were excluded as duplicates and papers on hearing loss, psychiatric disorders, trauma, Huntington disease, postpartum and pregnancy, papers on 36-item version, and general commentaries. The remaining 43 records were screened based on their titles and abstracts and 13 irrelevant papers were excluded. The number of observed agreements between the reviewers was 30 (70% of the observations) and kappa was 0.23 (SE 0.16, 95% CI -0.09 to 0.55) considering the strength of agreement between the reviewers to be 'fair'. Thirty records were assessed based on their full-texts and 16 irrelevant papers were excluded comprising 14 relevant studies potentially fit for a qualitative analysis (Figure 1). Additionally, one study that was published after the search was considered relevant into further analysis [6].

Data extraction

The attempt to extract relevant data regarding a 12-item self-administered WHODAS 2.0 version from the report by Tazaki et al. [7] was unsuccessful and that study was excluded from further analysis. Tazaki et al. [7] employed five different versions of WHODAS 2.0 (12- and 36-item interviewer-administered, 36-item proxy-administered, and 12- and 36-item self-administered versions) and there was a discrepancy in reporting a sample size (total n=126 but 62 men and 70 women). After the selection and data extraction phases, 14 records were included into further analysis.

Studied samples

All of the 14 remaining papers were published after 2013 (Table 2). All of them were observational studies. Two studies focused on the elderly [8, 9] while the rest evaluated people of working age. The sizes of samples varied from 80 up to 31,251 participants. Except for one study with 98% women [10], the proportions of female participants were between 47% and 65%. A great diversity was observed in the participants' health problems: patients waiting for an elective joint arthroplasty or neurosurgery [8, 11], general population or healthy volunteers [4, 9, 12, 13], patients with chronic musculoskeletal pain or fibromyalgia [10, 14, 15], patients with spinal cord injury [5, 16], and people reimbursed for any disabilities [17].

Psychometric properties

The most common psychometric properties reported by the included studies were Cronbach's alpha and convergent validity. The alpha estimates were usually high varying from 0.81 up to 0.96. Any pooling of the reported concurrent validity estimates was impossible as each study compared WHODAS 2.0 with different scales. However, the reported correlations between WHODAS 2.0 and other disability scales applied at the same time with WHODAS 2.0 were high in most of the studies. Floor and ceiling effects were reported by three studies. One study (the biggest sample size of 31,251) reported a substantial floor effects up to 32% (on average 20%) without a ceiling effect [12]. Another study conducted on a sample of 183 participants did not observe any floor or ceiling

effects [11]. Moreover, in that study, none of the participants – patients waiting for a neurosurgical procedure – reported a highest or lowest WHODAS 2.0 scores. The third study reported a significant floor effect up to 80% for all 12 items and for a total score without a ceiling effect [6].

Exploratory factor analysis or principal component analysis were employed by six studies [5, 10, 12, 14, 15, 17]. None of them reported a unidimensional structure of 12-item WHODAS 2.0. The number of factors varied from two up to five. Four studies employed confirmatory factor analysis. Only one of them reported a good model fit [4]. In one study, the hierarchical model with one common factor and six subfactors (as suggested by WHODAS 2.0 developers for 36-item version) was assessed resulting in poor fit [5]. One study reported a good fit of one-factor model but the reported root mean square error of approximation (RMSEA) was insignificant 0.079 pointing at a poor fit [11]. Another study reported a good fit of a two-factor model [15].

In one study, the discriminative ability was assessed using Karnofsky Performance Status scale as indicator of disability [11] reporting positive results. Another study assessed the discrimination ability using the item response theory [14]. That study reported discrimination of WHODAS 2.0 items being high to perfect, even though, the difficulty of items was shifted towards elevated disability rates. Such a shift implicates that a respondent should be experiencing slightly worse disability (compared with the average population rate) to achieve a 50/50 probability of giving an answer that would be interpreted by the WHODAS 2.0 as a "worse disability."

Three studies assessed test-retest reliability of the 12-item WHODAS 2.0 [9, 13, 18] reporting insignificant differences between repeated measures. In all three studies, the time interval between measures was one week.

DISCUSSION

This systematic review of 14 observational studies evaluated the available evidence on the psychometric properties of the self-administered 12-item WHODAS 2.0 among a general adult population or people with non-acute physical causes of disability. While the spectrum of the studies was expectedly wide, some patterns could be observed. Firstly, most of the studies found WHODAS 2.0 to be internally consistent. Secondly, the scale seemed reliable in term of test-retest reproducibility even if the time interval between studied repeated measures was hardly sufficient (one week). Thirdly, the 12-item WHODAS 2.0 might have a substantial floor but not ceiling effect. Therefore, the screening ability of this WHODAS 2.0 version seems to be weak as it may not distinguish lower levels of disability severity. Fourthly, WHODAS 2.0 seems to be able to discriminate well people with other than the lowest levels of perceived disability. Fifthly, respondents might be slightly more disabled in reality than the reported level of disability implies. Finally, the biggest concern risen of this review is one regarding the factor structure of WHODAS 2.0. Instead of unidimensionality, several included studies pointed at the multidimensional structure of the scale. While unidimensionality refers to measuring a single construct (in this case, disability level), multidimensionality refers to the fact, that scale is measuring two or several different constructs. That makes the total scores of multidimensional tests hard to interpret as there is no certainty on the exact contributions of each underlying construct to the total [19].

The main weakness of this systematic review was the considerable heterogeneity of the included papers. Their study populations ranged from healthy volunteers to tetraplegics. While one of the advantages of WHODAS is comparability between different health problems, conclusions could be more reliable if there were several studies on a similar disorder in different settings and on large samples. The number of identified relevant studies was surprisingly small. The included studies assessed convergent validity of WHODAS 2.0 by comparing with a wide spectrum of different tests and scales. While those comparators were mostly valid and reliable, the small number of studies on each of them made a reliable pooling impossible. Unfortunately, no system of the assessment of systematic bias seemed to fit the purpose of the review. The uncertainty regarding the methodological quality of the included studies may substantially weaken the strength of generalization of the results. This was, however, the first attempt to evaluate systematically the properties of the 12-item self-administered WHODAS 2.0 and the review was able to deliver several generalized clinical recommendations.

Only one previous review has been conducted on the topic so far [3]. Evaluating over 800 papers on the WHODAS 2.0, Federici et al. concluded that the WHODAS 2.0 shows strong correlations with several other measures of activity limitations probably due to the fact that it shares the same disability latent variable with them. This good convergent validity was in line with the findings of the present review. Concerning the factor structure of WHODAS 2.0, the conclusions of review by Federici et al. were more optimistic than the inferences of the present review that could not confirm the one-factor structure of 12-item WHODAS 2.0. The differences in the results of these two reviews may lay in the differences between their scopes. The scope of the present review was limited to a self-reported version of 12-item WHODAS 2.0 applied to a general population and people with non-acute physical

conditions. It is possible that the factor structures of other forms of WHODAS 2.0 are different. It is also possible that WHODAS 2.0 may behave differently when applied to populations others than studied here. It has to be noted that the majority of the papers included into the present study were published after April 2016 when the review by Federici et al. was already submitted.

This review focused on a self-reported version of WHODAS 2.0. The psychometric properties of interviewer- and proxy-administered forms may be different. When giving a self-reported response, a respondent may exaggerate, avoid embarrassing details, or try to confirm a guessed research question. A response may also be affected by the desire to obtain some social or financial benefit or service. On the other hand, a self-reported test may avoid the influence of interaction with an assessor.

Implications for clinical practice

Due to a substantial floor effect, the use of the self-administered 12-item WHODAS 2.0 as a screening tool in general population, when only mild severity of disability is expected, seems questionable. This scale may be used as an easy-to-use short questionnaire to assess the functioning profile of people with chronic physical conditions. The 12-item WHODAS 2.0 seems to be able to produce reliable repeated measures and, thus, may be used to assess the change in functioning level. Due to its multidimensional structure (measuring more than a single underlying construct), the 12-item version of WHODAS 2.0 may not be able to produce a reliable and comparable total score. Instead, the scale's 12 items should be scored and presented separately as a profile.

Recommendations for further research on 12-item WHODAS

The discrimination of this scale version ability is poorly understood – only two studies are conducted on the subject so far, each employing a different statistical technique [11, 14]. The minimal clinically important difference and minimal detectable change of the scale are still unknown and should be studied separately for each of the 12 items due to a seemingly certain multidimensional structure of the 12-item version. The convergent validity should be re-tested against similar relevant standard scales. The results of the item response theory obtained from only one sample should be reproduced in different settings and populations. The test-retest reliability assessment should be repeated in different time interval between test-retest measures. A short time interval (like a one-week interval employed in the included studies) may make the carryover effects due to memory, practice, or mood more probable. Instead, longer intervals increase the probability of changes in the clinical status [20, 21]. When a reference test (gold standard) is applicable then the sensitivity and specificity of WHODAS 2.0 should be evaluated, at least, in some populations.

Conclusions

It seems, that the 12-item self-administered WHODAS 2.0 is internally consistent and a reliable scale demonstrating overall good correlation with other measures of disability. However, it appears that it is a multidimensional scale and its total score may represent different combinations of several contributing factors.

Thus, the 12-item WHODAS 2.0 can be more reliable when creating a person's functional profile formed by the 12 individual item scores instead of a single total sum.

REFERENCES

[1] Üstün TB, Chatterji S, Kostanjsek N, Rehm J, Kennedy C, Epping-Jordan J, et al. Developing the World Health Organization Disability Assessment Schedule 2.0. Bull World Health Organ. 2010;88(11):815-23.

[2] World Health Organisation. WHO Disability Assessment Schedule 2.0 WHODAS 2.0, Psychometric Qualities: WHO 2014 [cited 2015 October 16]. Available from: www.who.int/classifications/icf/whodasii/en/index2.html.

[3] Federici S, Bracalenti M, Meloni F, Luciano JV. World Health Organization disability assessment schedule 2.0: An international systematic review. Disability and rehabilitation. 2017;39(23):2347-80.

[4] Kimber M, Rehm J, Ferro MA. Measurement Invariance of the WHODAS 2.0 in a Population-Based Sample of Youth. PloS one. 2015;10(11):e0142385.

[5] Smedema SM, Ruiz D, Mohr MJ. Psychometric Validation of the World Health Organization Disability Assessment Schedule 2.0-Twelve-Item Version in Persons With Spinal Cord Injuries. Rehabilitation Research Policy and Education. 2017;31(1):7-20.

[6] Katajapuu N, Laimi K, Heinonen A, Saltychev M. Floor and ceiling effects of the World Health Organization Disability Assessment Schedule 2.0 among patients with chronic musculoskeletal pain. International journal of rehabilitation research Internationale Zeitschrift fur Rehabilitationsforschung Revue internationale de recherches de readaptation. 2019.

[7] Tazaki M, Yamaguchi T, Yatsunami M, Nakane Y. Measuring functional health among the elderly: development of the Japanese version of the World Health Organization Disability Assessment Schedule II. International journal of rehabilitation research Internationale Zeitschrift fur Rehabilitationsforschung Revue internationale de recherches de readaptation. 2014;37(1):48-53.

[8] Galli T, Mirata P, Foglia E, Croce D, Porazzi E, Ferrario L, et al. A comparison between WHODAS 2.0 and Modified Barthel Index: which tool is more suitable for assessing the disability and the recovery rate in orthopedic rehabilitation? ClinicoEconomics and outcomes research : CEOR. 2018;10:301-7.

[9] Silva AG, Cerqueira M, Raquel Santos A, Ferreira C, Alvarelhao J, Queiros A. Inter-rater reliability, standard error of measurement and minimal detectable change of the 12-item WHODAS 2.0 and four performance tests in institutionalized ambulatory older adults. Disability and rehabilitation. 2017:1-8.

[10] Smedema SM, Yaghmaian RA, Ruiz D, Muller V, Umucu E, Chan F. Psychometric validation of the world health organization disability assessment schedule 2.0-12-item Version in persons with fibromyalgia syndrome. Journal of Rehabilitation. 2016;82(3):28-35.

[11] Schiavolin S, Ferroli P, Acerbi F, Brock S, Broggi M, Cusin A, et al. Disability in Italian neurosurgical patients: validity of the 12-item World Health Organization Disability Assessment Schedule. International journal of rehabilitation research Internationale Zeitschrift fur Rehabilitationsforschung Revue internationale de recherches de readaptation. 2014;37(3):267-70.

[12] Gaskin CJ, Lambert SD, Bowe SJ, Orellana L. Why sample selection matters in exploratory factor analysis: implications for the 12-item World Health Organization Disability Assessment Schedule 2.0. BMC medical research methodology. 2017;17(1):40.

[13] Marom BS, Carel RS, Sharabi M, Ratzon NZ. Cross-cultural adaptation of the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) for Hebrew-speaking subjects with and without hand injury. Disability and rehabilitation. 2017;39(12):1155-61.

[14] Saltychev M, Bärlund E, Mattie R, McCormick Z, Paltamaa J, Laimi K. A study of the psychometric properties of 12-item World Health Organization Disability Assessment Schedule 2.0 in a large population of people with chronic musculoskeletal pain. Clinical rehabilitation. 2017;31(2):262-72.

[15] Saltychev M, Mattie R, McCormick Z, Laimi K. Confirmatory factor analysis of 12-Item World Health Organization Disability Assessment Schedule in patients with musculoskeletal pain conditions. Clinical rehabilitation. 2017;31(5):702-9.

[16] Tarvonen-Schröder S, Kaljonen A, Laimi K. Utility of the World Health Organization Disability Assessment Schedule and the World Health Organization minimal generic set of domains of functioning and health in spinal cord injury. Journal of rehabilitation medicine. 2018;51(1):40-6.

[17] Xenouli G, Xenoulis K, Sarafis P, Niakas D, Alexopoulos EC. Validation of the World Health Organization Disability Assessment Schedule (WHO-DAS II) in Greek and its added value to the Short Form 36 (SF-36) in a sample of people with or without disabilities. Disability and Health Journal. 2016;9(3):518-23.

[18] Moreira A, Alvarelhão J, Silva AG, Costa R, Queirós A. Validation of a Portuguese version of WHODAS 2.0 - 12 items in people aged 55 or more. Revista Portuguesa de Saude Publica. 2015;33(2):179-82.

[19] Ravaud JF, Delcey M, Yelnik A. Construct validity of the functional independence measure (FIM): questioning the unidimensionality of the scale and the "value" of FIM scores. Scand J Rehabil Med. 1999;31(1):31-41.

[20] Brown G, Irving E, Keegan P. An Introduction to Educational Assessment, Measurement and Evaluation. 2 ed. Rosedale, North Shore, New Zealand: Pearson Education; 2008.

[21] Multon KD. Test–Retest Reliability 2012 [cited June 24, 2019]. In: Encyclopedia of Research Design [Internet]. Thousand Oaks, CA, USA: SAGE Publications, [cited June 24, 2019]; [2-5]. Available from: https://methods.sagepub.com/base/download/ReferenceEntry/encyc-of-research-design/n457.xml.

Table 1. Search strategy

Database	Search clauses and filters
Medline (PubMed)	(whodas [TI] OR "World Health Organization Disability Assessment Schedule" [TI] OR “who-das” [TI] OR “who das” [TI]) AND (“12” OR “twelve”) AND (hasabstract[text])
Embase	(whodas:ti OR "World Health Organization Disability Assessment Schedule":ti OR “who-das”:ti OR “who das”:ti) AND (“12” OR “twelve”)
Scopus	(ALL(("12" OR "twelve"))) AND (TITLE((whodas OR "World Health Organization Disability Assessment Schedule" OR "who-das" OR "who das"))) AND (LIMIT-TO(DOCTYPE,"ar") OR LIMIT-TO(DOCTYPE,"le") OR LIMIT-TO(DOCTYPE,"no")) AND (LIMIT-TO(SRCTYPE , "j"))
Web of Science	(TITLE: (((wholes OR "World Health Organization Disability Assessment Schedule") OR "who-das") OR "who das") AND ALL FIELDS: ("12" OR "twelve")) Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. Refined by: DOCUMENT TYPES: (ARTICLE)
PsycINFO	TI ((whodas OR "World Health Organization Disability Assessment Schedule" OR “who-das” OR “who das”)) AND TX ((“12” OR “twelve”)) Source type: Academic Journals

Table 2. Basic characteristics of the included studies

Author and year	Country	Settings, participants and main diagnoses	Sample size	Women	Age, mean (standard deviation), years	Psychometric properties
Galli 2018[8]	Italy	Patients hospitalized for elective hip (60%) or knee (40%) arthroplasty (3 hospitals). After-surgery estimates excluded.	80	67%	70.1 (1.1)	Convergent validity with modified Barthel index: 0.335
Gaskin 2017[12]	Australia	SAGA data[1] - general population >=50 years	31,251	54%	63.4 (9.5)	EFA[5]: 1 to 3 factors (mostly 2 or 3). Floor effect: 6% to 32% (overall 20%). Ceiling effect: none.
Kimber 2015[4]	Canada	CCHS-MH data[2] - general population >=15 years youth group excluded.	23,798	51%	47.1 (0.2)	Alpha: 0.95 (95% CI 0.94 to 0.96) CFA[6] (assuming 1/6-factor structure).
Marom 2017[13]	Israel	Volunteers - general working population. Group with acute trauma excluded.	155	51%	43.1 (15.0)	Alpha: 0.85. Reported including patients with acute trauma: Test/retest (1 week): ICC[7] 0.88 (95% CI 0.83 to 0.91). Convergent validity: PCS-12[8] -0.46 (95% CI -0.67 to -0.15), MCS-15[9] -0.62 (95% CI -0.78 to -0.36), QDASH[10] 0.53 (95% CI 0.33 to 0.69).
Moreira 2015[18]	Portugal	General population using community support services.	144	64%	64.0 (6.7)	Alpha: 0.86. Test/retest (1 week): ICC 0.77(95% CI 0.69 to 0.83). Convergent validity: Barthel index: -0.27, LSNS[11] -0.19.
Saltychev 2017[14][3]	Finland	University outpatient clinic. Chronic non-specific musculoskeletal pain.	501	65%	47.1 (13.9)	EFA: 2 factors. IRT[12]: discrimination - high to perfect for all items difficulty - a slight shift towards elevated disability rates.
Saltychev2017[15] [4]	Finland	University outpatient clinic. Chronic non-specific musculoskeletal pain.	408	65%	47.0 (13.7)	EFA: 2 factors. CFA: 2-factor assumption.
Schiavolin 2014[11]	Italy	Patients scheduled for different neurosurgical surgery.	183	50%	51.1 (13.1)	CFA: assuming 1-factor structure (insignificant RMSEA 0.079). Alpha 0.875. Convergent validity: EUROHIS-QOL[13] -0.52, PGWBI-S[14] -0.52. Discriminative validity: significant difference between KPS[15]>90 and KPS=<90. Floor and ceiling effects: none (0%).
Silva 2017[9]	Portugal	Day Care Centers and Nursing Homes. Possibly including interviewer-administered version of WHODAS 2.0.	100	62%	82.3 (8.1)	Convergent validity: GST[16] –0.57 to –0.62, FTSTS[17] 0.41, TUG[18] 0.32 to 0.37. Test/retest (1 week): p-value 0.32
Smedema 2017[5]	USA	Online survey. Patients with spinal cord injury.	247	50%	41.6 (12.4)	Alpha: 0.82. CFA: hierarchical and 1-factor with poor fit. EFA: 3 factors. Convergent validity was reported for each of 3 factors separately. Convergent validity: SWLS[19] -0.16 to -0.36, CSES[20] -0.05 to -0.56, IPA[21] -0.16 to -0.47, SF-20[22] 0.0 to -0.62.
Smedema 2016[10]	USA	Online survey. Patients with self-reported fibromyalgia.	302	98%	48.4 (10.4)	PCA[23]: 2 factors. 1st factor: alpha: 0.81; convergent validity: BFI[24] 0.35, pain intensity 0.28, MOS-Sleep[25] 0.33, GFQ[26] 0.53, CESD-10[27] 0.63, MSPSS[28] -0.37. 2nd factor: alpha 0.83; convergent validity BFI 0.43, pain intensity 0.43, MOS-Sleep 0.43, CFQ 0.31, CESD-10 0.42, MSPSS -0.20.
Tarvonen-Schröder 2018[16]	Finland	University outpatient clinic. Patients with spinal cord injury.	142	47%	56.7 (16.9)	Alpha: 0.86. Convergent validity:7-item World Health Organization (WHO) minimal generic set 0.49.
Xenouli 2016[17]	Greece	People without (A) and with (B) disabilities	109/ 101	65% / 63%	46.3 (13.0) / 51.5 (18.4)	EFA: group A – 5 factors, group B – 4 factors. Groups A + B: Alpha 0.85. Convergent validity: SF-PCS[29] -0.76, SF-MCS[30] -0.50, PSS-14[31] 0.55
Katajapuu 2019 [6] [5]	Finland	University outpatient clinic. Chronic non-specific musculoskeletal pain.	1988	65%	47.6 (15.0)	Floor effect: 15% to 79%. Ceiling effect: none.

[1] World Health Organization's longitudinal Study on global ageing and adult health (6 countries); [2] Canadian Community Health Survey-Mental Health; [3] Subpopulation of Katajapuu 2019 [6]; [4] Subpopulation of Katajapuu 2019 [6]; [5] Exploratory factor analysis; [6] Confirmatory factor analysis; [7] Intraclass correlation coefficient; [8] Physical composite scores; [9] Mental composite scores; [10] Quick Disability of Arm, Shoulder, and Hand Outcome Measure; [11] Lubben Social Network Scale; [12] Item response theory analysis; [13] European Health Interview Survey-Quality of Life; [14] Psychological General Well-Being Index-Short; [15] Karnofsky Performance Status; [16] Gait speed test; [17] FTSST: Five-times-sit-to-stand-test; [18] Time Up & Go Test; [19] Satisfaction with Life Scale; [20] Core Self-Evaluations Scale; [21] Work and Education subscale of the Impact on Participation and Autonomy

Questionnaire; [22] Medical Outcomes Study 20-Item Short-Form Health Survey; [23] Principal Component Analysis; [24] Brief Fatigue Inventory; [25] Medical Outcomes Study – Sleep Scale; [26] Cognitive Failures Questionnaire; [27] Center for Epidemiological Studies Depression Scale-Short Form; [28] Multidimensional Scale of Perceived Social Support; [29] Short Form 36 Brief Physical Health Scale; [30] Short Form 36 Brief Mental Health Scale; [31] Perceived Stress Scale

Figure 1. Search flow