JYU DISSERTATIONS 368

Niina Katajapuu

Psychometric Properties of 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) Amongst People with Chronic Musculoskeletal Pain

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF SPORT AND
HEALTH SCIENCES

JYU DISSERTATIONS 368

Niina Katajapuu

Psychometric Properties of 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) Amongst People with Chronic Musculoskeletal Pain

Esitetään Jyväskylän yliopiston liikuntatieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi Liikuntarakennuksen auditoriossa L304
huhtikuun 16. päivänä 2021 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Sport and Health Sciences of the University of Jyväskylä,
in building Liikuntarakennus, auditorium L304 on April 16, 2021 at 12 o'clock..

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2021

Editors
Anne Viljanen
Faculty of Sport and Health Sciences, University of Jyväskylä
Päivi Vuorio
Open Science Centre, University of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-8587-5

ISBN 978-951-39-8587-5 (PDF)
URN:ISBN:978-951-39-8587-5
ISSN 2489-9003

ABSTRACT

Katajapuu, Niina
Psychometric Properties of 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) Amongst People with Chronic Musculoskeletal Pain
Jyväskylä: University of Jyväskylä, 2021, 71 p.
(JYU Dissertations
ISSN 2489-9003; 368)
ISBN 978-951-39-8587-5 (PDF)

Aim of this thesis was to explore the floor and ceiling effect, differential item functioning (DIF), minimal clinically important difference (MCID) and minimal detectable change (MDC) of 12-item World Health Organization Disability Assessment Schedule (WHODAS) 2.0 amongst people with chronic musculoskeletal pain. Of the cross-sectional data of 1 988 patients seen at Physical and Rehabilitation Medicine (PRM) clinic, 65 % were women. The mean age was 48 years, and the average score of 12–item WHODAS 2.0 was 13/48 points. Of the participants, 88% had a primary diagnosis of 'Diseases of the musculoskeletal system and connective tissue'. Of those, 39% had a diagnosis of 'Dorsalgia'. Floor and ceiling effects were calculated as relative frequencies of the lowest or the highest scores for each item. The DIF was tested using logistic regression analysis and the Chi^2 test and interpreted graphically based on item characteristic curves. The MDC and MCID were calculated based on the variance of the scores. A systematic review explored the evidence of the psychometric properties of 12-item WHODAS 2.0 among general population and people with non-acute physical causes of disability.

The 12-item WHODAS 2.0 demonstrated high, almost nine points MDC. As the MDC exceeded the level of MCID, nine points were considered to be the amount of change perceived by a respondent as clinically significant. A significant floor effect (>15%) was seen in all 12 items. A significant uniform gender-related DIF was detected in 7 of 12 items.

In conclusion, due to the floor effect, the 12-item WHODAS may have limitations in the lower end of the scale amongst people with milder disability. Seven items functioned differently between men and women and almost nine points MDC might complicate the use of WHODAS 2.0 total score. It appears that it is a multidimensional scale, and its total score may represent different combinations of several contributing factors. Therefore, all these findings should be taken into consideration while making the work ability or rehabilitation evaluations or interpreting the results based on the single total score instead of item scores in patients with chronic musculoskeletal pain.

Keywords: WHODAS 2.0, psychometrics, rehabilitation, musculoskeletal disease, pain

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Katajapuu, Niina
12-osaisen WHODAS 2.0 -mittarin psykometriset ominaisuudet kroonisilla tuki- ja liikuntaelinkipupotilailla
Jyväskylä: Jyväskylän yliopisto, 2021, 71 s.
(JYU Dissertations
ISSN 2489-9003; 368)
ISBN 978-951-39-8587-5 (PDF)

Tämä väitöskirjatutkimus selvitti 12-osaisen terveyden ja toimintarajoitteiden arviointimenetelmän WHODAS 2.0 -mittarin toimivuutta kroonisilla tuki– ja liikuntaelinkipupotilailla. Mittarin katto– ja lattiavaikutusta, mittarin osioiden muuttumattomuutta (DIF) sekä mittarin pienintä havaittavaa muutosta (MDC) ja pienintä kliinisesti tärkeää eroa (MCID) selvitettiin fysiatrian poliklinikalta poikkileikkausasetelmassa kerätyllä (N=1988) aineistolla. Tutkimusjoukko koostui naisista (65%) ja miehistä, joiden keskiarvoikä oli 48 vuotta. Mittarin kokonaispistemäärän keskiarvo oli 13/48. Tutkittavista 88 %:lla oli päädiagnoosi 'Tuki- ja liikuntaelimistön sairaus' ja heistä 39 %:lla 'Selkäkipu'. Katto– ja lattiavaikutus laskettiin vastaajien kokonais- ja osiopistemäärien frekvensseistä. Mittarin DIF tarkasteltiin logistisella regressioanalyysillä sekä Chi^2 testillä. Osioiden toimivuutta tarkasteltiin kahden parametrin osiovaste teorialla. MDC ja MCID laskettiin perustuen pistemäärän varianssiin. Systemaattisella kirjallisuuskatsauksella selvitettiin tietoa 12-osaisen WHODAS 2.0 -mittarin psykometrisista ominaisuuksista normaaliväestöllä ja ihmisillä, joilla oli fyysisiä sairauksia ja toimintarajoitteita. 12 osion WHODAS 2.0 -mittarin MDC oli 8,6 pistettä ylittäessä MCID: n kynnysarvon, joten vasta 9 pisteen muutos koetaan käytännössä toimintakyvyn muutokseksi. Mittarin kaikissa osioissa havaittiin tilastollisesti merkitsevä lattiaefekti (>15 %) ja erilainen osioiden toimivuus miesten ja naisten välillä 7:ssä osiossa.

Johtopäätökset: lattiaefektistä johtuen mittarin erottelukyky asteikon matalammassa päässä saattaa olla huono lievissä toimintakyvyn rajoituksissa. Seitsemän mittarin osiota toimii eri tavalla miehillä ja naisilla. Lisäksi huomattavan korkea pienin havaittu muutos saattaa vaikeuttaa mittarin antaman kokonaispistemäärän käyttämistä ja tulosten tulkintaa kroonisilla tuki- ja liikuntaelinkipupotilailla. Näyttää siltä, että 12 osion WHODAS 2.0 on moniulotteinen mittari, jonka kokonaispistemäärän muodostumiseen vaikuttaa usea tekijä. Hyödyntäessä mittarin kokonaispistemäärää osiopistemäärän sijaan kroonisten tuki- ja liikuntaelinkipupotilaiden työkykyarviossa ja kuntoutusintervention vaikutusten arvioinnissa, edellä kuvatut löydökset tulee ottaa huomioon.

Avainsanat: WHODAS 2.0, psykometriikka, kuntoutus, tuki- ja liikuntaelinsairaudet, kipu

Author

Niina Katajapuu, PT, MSc
Faculty of Sport and Health Sciences
University of Jyväskylä
Finland
niina.katajapuu@turkuamk.fi
ORCID 0000-0002-7416-8928

Supervisors

Professor Ari Heinonen, PhD
Faculty of Sport and Health Sciences
University of Jyväskylä
Finland

Professor Mikhail Saltychev, MD, PhD
Department of Physical and Rehabilitation Medicine
Department of Clinical Medicine
University of Turku
Finland

Reviewers

Senior Researcher Heidi Anttila, PhD
Department of Welfare
Finnish Institute for Health and Welfare
Finland

Associate Professor Markku Kankaanpää, MD, PhD
Department of Physical and Rehabilitation Medicine
University of Tampere
Finland

Opponent

Professor Anne Söderlund, PhD
School of Health, Care and Social Welfare
Mälardalen University
Sweden

ACKNOWLEDGEMENTS

For everything there is a fixed time, and a time for every business under the sun"Kaikella on määrähetkensä, aikansa joka asialla taivaan alla" (Ecclesiastes / Saarnaaja 3:1)

The University of Jyväskylä has been my academic home since 1988. My first touch on this exhilarating environment was when I met the university students in my high school years in Liikunta building completing their teaching practise in my physical education class. Ever since, I decided to join this community. Years later I finished my master studies in the same building and today, with joy and gratefulness for the University of Jyväskylä, I will finalize my studies in this same academic home.

I want to express my gratitude to the reviewers, Senior Researcher Heidi Anttila and Associate Professor Markku Kankaanpää for going deep into this work and giving your expertise to make this thesis even better. I also feel truly privileged to have Professor Anne Söderlund as my opponent on this very special day of public defence.

I really want to give my warm appreciation to my both supervisors. Professor Ari Heinonen, you have given me the possibility to learn both, the independent research work and at the same time to be part of the randomized controlled study research group in the faculty. These two and half years have been the most instructive years in my academic life. And Professor Mikhail Saltychev. First, I want to thank you from my heart for turning my head from my original research idea on the other and second, for giving me the opportunity to utilize this data from the Turku University Hospital for this dissertation. I also want to thank you for answering my endless questions or not answering, but pushing me to learn the surface of the difficult field of psychometrics. Thank you for your support and encouraging supervision through these years.

I also want to show my gratitude to my closest Liikunta colleagues. Eeva Aartolahti, Maarit Janhunen and Outi Ilves, thank you for giving me the best formal and informal academic tips and laughs during these couple of years. I also want to thank all my colleagues in Turku University of Applied Sciences, especially Kati Kulju for making me rewrite my Finnish summary and Marion Karppi for skiing with me my brains out in "Vale Riviera" while waiting for the last feedback from the reviewers. Thank you Ursula Hyrkkänen for all the Wau-words you have said after each step I reached during this academic trip. My great gratitude goes also to my friend Chrisna Ravyse. You are one of the bravest and smartest woman I know in this world. Thank you. I also want to thank Pirjo Pennanen for just being there from the very beginning.

Because life is not just writing and studying, I also want to thank all my friends, especially Mervi and Johanna and all the Tri Babes around the Finland for swim, bike and run and all the coffee and donuts during this wonderful trip. Thank you for being there for me and reminding me of the real life.

My most emotional gratefulness goes for my parents Erkki and Tiina and my sisters Piia and Lotta, my “cousinsister” Sanna-Maija and all the “Peetun kamut” group around the world. You have prepared me for this moment since I was born. I want to thank you for your real interest on whatever I have engaged in my life. No matter how big or small my crazes have been, your interest, proudness and joy has been real.

And last, I want to thank these three men in my life who mean everything to me. Kasper and Oskar and Esa. Thank for every each of you on your own way making me to understand, that “For everything there is a fixed time, and a time for every business under the sun” (Ecclesiastes 3:1)

Turku February 21.3.2021

Niina Katajapuu

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, which are referred in the text by Roman numerals:

I Saltychev M, Katajapuu N, Bärlund E, Laimi K. 2019. 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 vol 23, 1-6
DOI: 10.1080/09638288.2019.1643416

II Katajapuu N, Laimi K, Heinonen A, Saltychev M. 2019. 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 vol 42, 190-192
DOI: 10.1097/MRR.0000000000000339

III Katajapuu N, Laimi K, Heinonen A, Saltychev M. 2019. Gender-related differences in psychometric properties of WHO Disability Assessment Schedule 2.0. International Journal of Rehabilitation Research vol 42, 316-321
DOI: 10.1097/MRR.0000000000000365

IV Katajapuu N, Heinonen A, Saltychev M. 2020. Minimal clinically important difference and minimal detectable change of the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) amongst patients with chronic musculoskeletal pain. Clinical Rehabilitation vol 34, 1506-1511
DOI: 10.1177/0269215520942573

The author has substantially contributed to the conception and design, analysis, writing and interpretation of the results of all four publications. The author has been responsible for drafting the works II-IV and has an equal contribution to the data collection and writing with the first author for the work I. The author has approved the final versions of all four works to be published. The author is accountable for all aspects of all four publications in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FIGURES

TABLES

ABBREVIATIONS

BMI	Body Mass Index
CTT	Classical Test Theory
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
ICD	International Classification of Diseases
ICIDH	International Classification of Impairments, Disabilities and Handicaps
ICF	International Classification of Functioning, Disability and Health
IRT	Item Response Theory
KELA	Social Insurance Institution of Finland [Kansaneläkelaitos]
MCID	Minimal Clinically Important Difference
MDC	Minimal Detectable Change
MDS	Model Disability Survey
PRM	Physical Rehabilitation Medicine
PROMIS	Patient Reported Outcome Measure Information System
SD	Standard Deviation
THL	Finnish Institute for Health and Welfare [Terveyden ja hyvinvoinnin laitos]
UN	United Nations
WHO	World Health Organization
WHODAS 2.0	World Health Organization Disability Assessment Schedule
WHOQL	World Health Organization Quality of Life scale

CONTENTS

1 INTRODUCTION

The Finnish Ministry of Social Affairs and Health set the specialist board for the rehabilitation reform in 2016. Based on the board's 55 proposals, the experts of the Rehabilitation Knowledge Base project ['Kuntoutuksen tietopohja –hanke'] published in May 2020, the new guidelines for the self-assessment of functioning. These guidelines are intended for use amongst adults. The purpose is to help to recognize their need for rehabilitation and to measure the outcomes of the rehabilitation. The guidelines suggest three generic self-evaluation measures to be used to evaluate functioning (Sosiaali - ja terveysministeriö 2017) (Ministry of Social Affairs and Health, 2017). One of these three measures is the 12-item World Health Organisation Disability Assessment Schedule (WHODAS 2.0), which is based on the International Classification of Disability, Functioning and Health (ICF) framework (WHO 2001). The Finnish legislation requires that functioning must be assessed before granting rehabilitation or other comparable services to a citizen (THL 2019a). The Social Insurance Institution of Finland (Kela) demands from the providers of rehabilitation services using the ICF framework when evaluating clients' functioning (Kela 2020c). The Kela provides a detailed list of obligatory outcome measures for musculoskeletal rehabilitation courses. The list does not include any particular measure of the comprehensive overall functioning level. Instead, service providers are guided to use a national TOIMIA network and its database of outcome measures (Kela 2020a, Kela 2020b).

The TOIMIA network provides national guidelines for evaluating functioning using the ICF framework. These guidelines proposed using the WHODAS 2.0 to evaluate participation amongst adults, especially those who are suffering from physical disabilities (Paltamaa & Kantanen 2013). The TOIMIA is describing psychometric properties of different scales (THL 2019c). Among other scales, the TOIMIA has extensively reviewed the psychometric properties of WHODAS 2.0 (Paltamaa & Anttila 2015). For that task, the available data concerning two versions of the scale of interest were evaluated: first – named "WHO-DAS II" or "WHODAS II" and the second version – the WHODAS 2.0. The last updated description of the WHODAS 2.0 presented by the TOIMIA includes the reliability and validity of the 36–item WHODAS 2.0. The validity of WHODAS 2.0 evaluation is described in the WHODAS manual (Ustun et al. 2010b), the WHO's World Multi-country survey (Ustun et al. 2003a), the World Health Survey (Ustun et al. 2003c) the Italian validation of WHODAS II by Federici et al. (2009 a) and in the review by Federici et al. (2009 b). Furthermore, the TOIMIA conducted a supplemental review in 2010-2014. In addition to these works, the description of the WHODAS 2.0 reliability was supplemented by the data extracted from numerous individual studies.

At least seven different name versions of the WHODAS can be found from the literature. Now, seven official versions of WHODAS 2.0, depending on the length and administration, exists (Ustun et al. 2010a). The majority of previous studies have been conducted using a 36-item version of the WHODAS 2.0. As regards the 12-item version of WHODAS 2.0, several psychometric properties and patient groups, to whom the scale has been applied, have been described. According to the paper by Ustun et al. (2010a), the 12-item version was able to explain 81% of the variability within the 36-item version (Ustun et al. 2010a). Amongst patients with severe depression, the measure is unidimensional with good discrimination ability, without differential item functioning between genders, and with good correlation with scales measuring the quality of life and the severity of depression (Luciano et al. 2010a, Luciano et al. 2010b, Luciano et al. 2010c). Amongst the elderly, the 12-item WHODAS 2.0 has also been found to be unidimensional (Sousa et al. 2010). The scale has been found unidimensional and without a differential, item functioning between genders amongst people with acute myocardial infarction (Kirchberger et al. 2014).

The overall knowledge on the psychometric properties of the 12-item WHODAS 2.0 has been scarce. Only a few validation studies of the WHODAS 2.0 Finnish translation has been published (Saltychev et al. 2017a, Saltychev et al. 2017b, Tarvonen-Schröder et al. 2018). There is a clear need for additional research concerning the psychometric properties of 12-item WHODAS 2.0 amongst people with musculoskeletal problems. The purpose of the present study was to evaluate the psychometric properties of 12-item self-administered WHODAS 2.0 amongst adults with chronic musculoskeletal pain. Such knowledge may help to implement the WHODAS 2.0 when recognizing the need for rehabilitation services, planning rehabilitation contents, and assessing achieved outcomes. If the 12-item WHODAS 2.0 turns out to be valid, it can be recommended for the nation-wide use as a fast and relatively simple measure of the functional level.

2 REVIEW OF THE LITERATURE

2.1 Functioning and disability

2.1.1 Classifications, models and definitions

World Health Organization (WHO) disseminates disease and disability classifications. Probably the most known is the International Classification of Diseases (ICD) published already in 1893. At the moment, the 11th revision (ICD-11) is ready to use. Classifications are needed to evaluate the health and health care process outcomes. ICD-10 is capable of evaluating the linear presence of impairment or disease, ignoring their consequences on activities or participation. One of the improvements from ICD-10 to ICD-11 is the supplemental material for functioning assessment including 36-item WHODAS 2.0 and Model Disability Survey (MDS) (WHO 1980, p. 10, Gray & Hendershot 2000, WHO 2019).

Forty years ago, the WHO's International Classification of Impairments, Disabilities and Handicaps (ICIDH) defined disability as a linear result of the impairment of the body structure or function. From that biomedical angle, a person might be restricted to perform every day activities by a disease only, and disability is a consequence of medical abnormality (WHO 1980, p. 11, Federici et al. 2009). Later development of ICIDH through ICIDH-2 resulted in ICF classification.

In ICF, the concepts disability and functioning are presented in the health-related functioning model (figure 1) (WHO 2001, p. 18). Compared to ICIDH, ICF model aims to describe the health-related functioning in biopsychosocial perspective, including the environmental and personal factors in the classification (WHO 2001, Salvador-Carulla & Garcia-Gutierrez 2011).

United Nations (UN) convention on the rights of persons with disabilities specifies disability based on its long-term nature, not as temporary or varying health state of a human being. In the UN convention (2006), the word disability is not that clearly defined. Instead, they describe the person *with* a disability or impairment. The Conventions definition lays more on the medical perspective, ignoring if the person's ability to participate in his life activities is decreased or not. From a human rights perspective, disability is defined as "results from the interaction between persons with (long-term) impairments and attitudinal and environmental barriers that hinders their full and effective participation in society on an equal basis with others”(Leonardi et al. 2006, United Nations 2006).

In Finland, Finnish Institute for Health and Welfare’s (THL) definition is based on the biopsychosocial model. Functioning includes "physical, psychological and social angles in peoples condition to cope with meaningful and necessary life activities in his environment" (THL 2019b). Definitions of the dimensions of functioning are presented in the next chapters.

Physical functioning

Physical functioning is related to optimal and safe mobility ability, what person has to have to be able to participate and act in varying life activities at home and other living environments (Satariano et al. 2012). The ICF classification (2001) presents that physical performance and physical capacity are near related to physical functioning. Physical performance describes an individual's ability to perform certain functions on the specific environmental situation, in turn, physical capacity relates on maximum capability on performing that function in standardized circumstances (WHO 2001, p. 14-15).

Mental and cognitive functioning

In Aalto et al. (2011) work, the dimensions of mental functioning include perception, thinking, learning and memory, which are needed to make representations of the environment and oneself. It consists of the ability to feel and experience the surrounding world. Mental functioning describes how the person can utilize his resources and abilities in different life situations and for the future. Like physical functioning, also mental functioning can be defined as maximal possible functioning or mental functioning in real-life situations (Aalto 2011, p. 1-2). Evans et al. (2000) claim in their work that to measure the mental and cognitive functioning, hundreds of tools have been developed, mostly based on the medical perspective of mental health. In general, the domains well-being, social functioning, problems or symptoms and risks to self or other people are covered in mental functioning when measuring the effectiveness of different therapies (Evans et al. 2000). Cognitive functions are part of the mental functions related to receive, manage, store and use the information and knowledge. Cognitive functioning is affected by different factors as mental alertness, circadian rhythm, mood and stress as an example (Tuulio-Henriksson 2011, p. 1).

Social functioning

Tiikkainen (2018) defines social functioning through two aspects: It includes social interaction and person as active participator and actor in society and organizations. Social functioning requires, e.g. the social skills and ability to interact and act in different roles. Social functioning requires the ability to adapt to the surrounding society with or without others, to participate in leisure time activities and to help each other (Tiikkainen & Pynnönen 2018, p. 1).

2.1.2 International Classification of Functioning, Disability and Health (ICF)

WHO (2001) developed the ICF classification, a renewal of International Classification of Impairments, Disabilities and Handicaps (ICIDH) to expand ICIDH's disease originated perspective towards a classification of health components (WHO 2001, p. 3-4). Before the final version of ICF, the ICIDH-2 was launched. It recognizes two relevant factors to understand and study disability and functioning; environmental factor and dimension of social participation (Gray & Hendershot 2000). In 2001 WHO accepted ICF classification, which aims to describe the interaction between health status, body structures and functions, activities, participation, environmental and individual factors concerning functioning or disability (WHO 2001, p. 18-19). Table 1 describes the concepts of ICF structure and figure 1 modified from WHO (2001, 18) illustrates the interaction between health status and ICF components. Compared to causal "process-like" ICIDH model, the multifaceted ICF model is more dynamic, and interaction flows to several directions affecting inhibitory or exhibitory on different functioning dimensions.

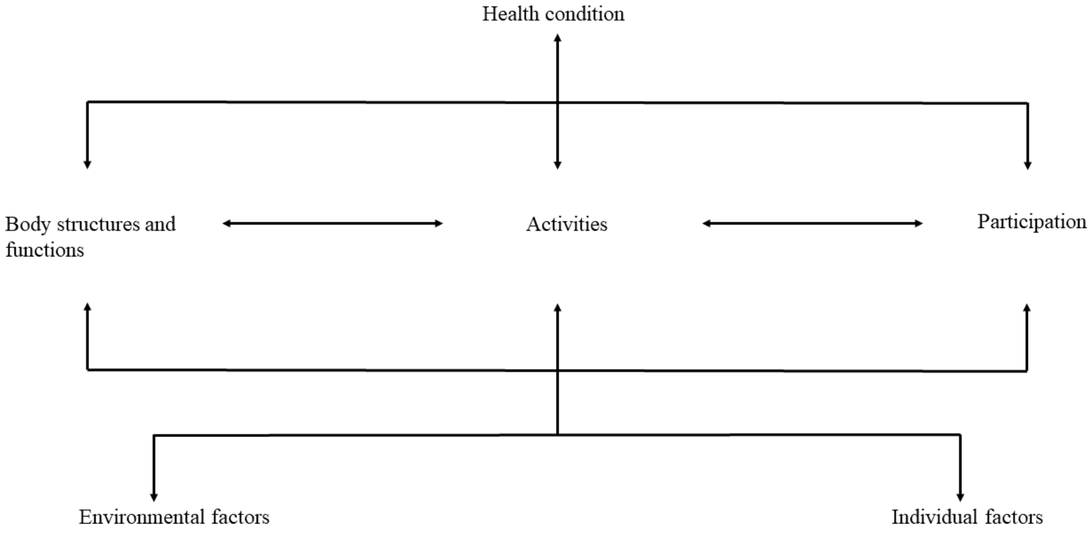


FIGURE 1 Interaction between the different ICF components

TABLE 1 ICF concepts

	Part 1 Functioning and disabilities		Part 2 Contextual factors	
Components	Body structures and functions	Activities and participation	Environmental factors	Individual factors
Domains	Body structures and functions	Life actions and tasks	External factors Influencing on functioning and disabilities	Internal factors influencing on functioning and disabilities

One use of the ICF classification is to give a common understanding of the concepts when measuring the disability or functioning. Several systematic reviews have been published to show the existing literature combining the disease-specific measurement tools and their linkage to ICF's four components (Wasiak et al. 2011, Oliveira et al. 2013, Naughton & Algar 2019). In WHO (2001) classification, the components 'Body structures and functions', 'Activities and participation' and 'Environmental factors' are divided into five to nine chapters each. Each chapter contains numerous codes. To be meaningful, the codes should contain at least one qualifier denoting the level of health on code in question. The restrictions on each code under each three components can be quantified using numeric scale from 0 denoting 'no problem', 1 'mild problem', 2 'moderate problem', 3 'severe problem', 4 'complete problem', 8 'not specified' and 9 'not applicable' (WHO 2001, p. 21-24).

ICF checklist

The comprehensive ICF classification consists of more than 1400 categories and their codes, which makes the full classification clinical use complex (Ptyushkin et al. 2012, p. 14). The ICF checklist (WHO 2003) was developed to solve the complexity of the ICF classification. ICF checklist includes the major ICF categories and most regularly used domains with the information of persons diagnose and other individual records. The qualifiers with the domains are used in the checklist to record the functioning and disability using various information sources. However, the 12-page checklist has over 120 domains to choose and qualify for generic assessment. Therefore the checklist has been used in ICF Core Set development (Stucki et al. 2008, Kostanjsek 2011).

ICF Core Set

To make ICF easier for everyday use, ICF Research Branch developed ICF Core Sets which should be used simultaneously with the International Classification of Diseases (ICD). ICF Core Sets are practical batteries of ICF categories and domains linked on different acute, post-acute and long-term health conditions. The development was aimed for clinical use, research work and for service providers to document and report functioning and disabilities. Now, ICF Research Branch has published seven ICF Core Set subcategories including musculoskeletal conditions, neurological conditions, mental health, other health conditions, diverse situations, cardiovascular and respiratory conditions and cancer. Though, the Core Sets are still missing the possibility to use them as self-reported measures, WHO worked further towards patient-reported measure, WHODAS questionnaire (Stucki et al. 2008, Ptyushkin et al. 2012, p. 19-21, ICF Research Branch 2017).

2.2 Background of the World Health Organization Disability Assessment Schedule WHODAS 2.0

Developing a generic assessment tool to evaluate functioning and disability is a long term process. Preceding the WHODAS 2.0, WHO coordinated several studies to develop a tool to measure psychiatric patients social functioning and to assess disabilities concerning mental disorders (Jablensky et al. 1980, WHO 1988, p. 77). First version of World Health Organization Disability Assessment Schedule (DAS) was followed by DAS II and D.A.S III as all the original items did not apply on outpatient population or the items were not found culturally relevant (Thara et al. 1988). The first version of WHO DAS theoretical framework was based on biomedical perspective where the disability and functioning were seen as a result of disease and the evaluation was focusing on body impairments and activities (Jablensky et al. 1980). WHODAS development proceeded almost parallel with the development of the ICIDH model towards ICF. On its own, ICF is not a practical measurement tool. Instead, the ICF framework created a shared understanding and a language of health-related functioning and disability. The present WHODAS 2.0 development was followed by WHODAS II. Both measures are based on the ICF framework to evaluate activity limitations and restrictions in participation (Prieto et al. 2000, Ustun et al. 2010a). For the development of WHODAS 2.0, information of disability models used in the health assessment questionnaires were used (Ziebland et al. 1993). Developers also utilized the information from the extensive project report "Measuring consumer outcomes in mental health" (Stedman 1997, p. 11-21). The parallel development of the disease and disability classifications, models and measure are presented in figure 2.

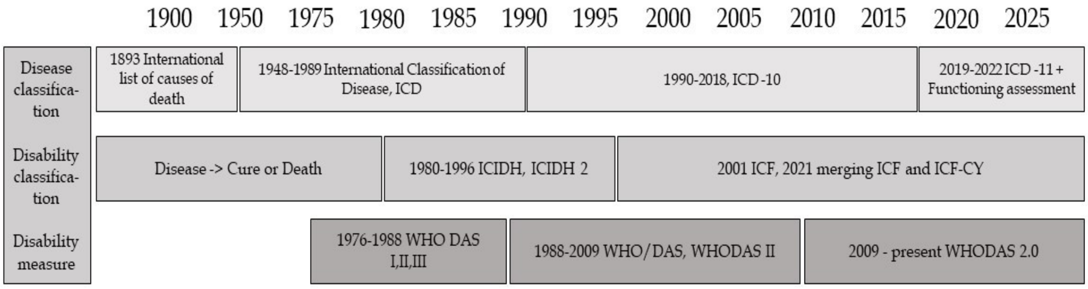

1900
1950
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
Disease classifica-tion
Disability classifica-tion
Disability measure
1893 International list of causes of death
1948-1989 International Classification of Disease, ICD
1990-2018, ICD-10
2019-2022 ICD -11 + Functioning assessment
Disease -> Cure or Death
1980-1996 ICIDH, ICIDH 2
2001 ICF, 2021 merging ICF and ICF-CY
1976-1988 WHO DAS I,II,III
1988-2009 WHO/DAS, WHODAS II
2009 - present WHODAS 2.0

FIGURE 2 Timeline for WHO disease and disability classifications and measure

2.3 Development of a measurement tool

Building a measurement tool has several steps, which are described in upcoming chapters reflecting the work of de Vet et al. (2011). Figure 3 summarizes the different phases and the contents of the measure development. The last and the continuous phase of the measure development, psychometric evaluation, is described in general level referring also to Mokkink et al. (2010) taxonomy. A detailed description of developing the WHODAS 2.0 and the existing literature of its psychometric properties will be described in chapter 2.4 and 2.5.

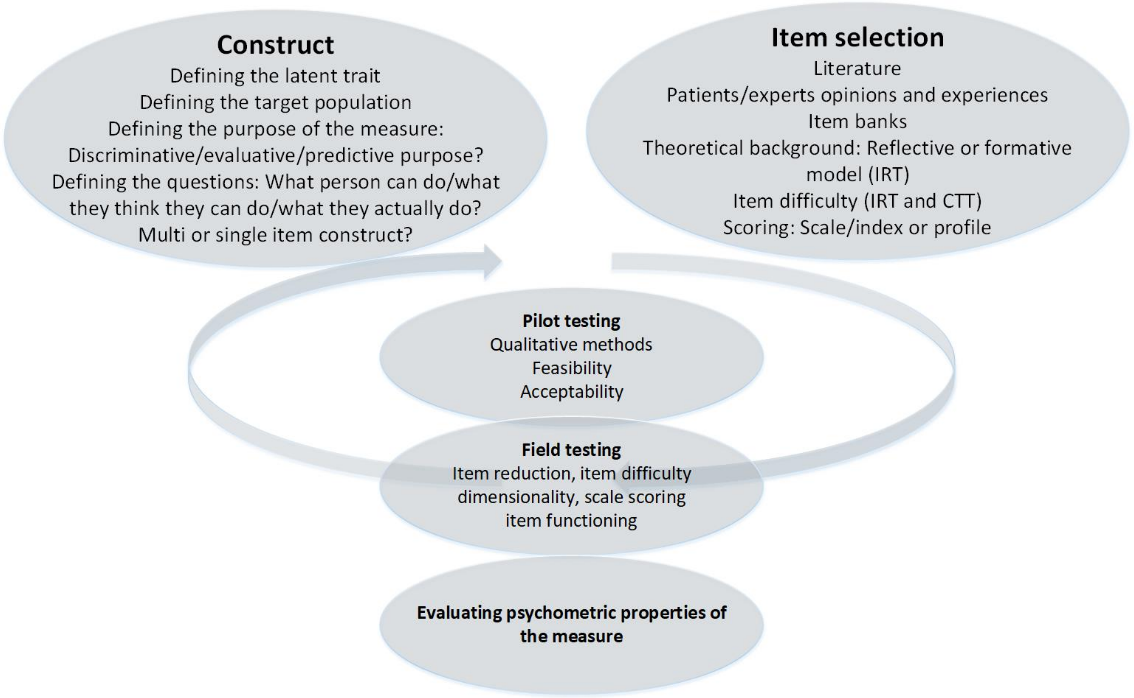

Construct
Defining the latent trait
Defining the target population
Defining the purpose of the measure:
Discriminative/evaluative/predictive purpose?
Defining the questions: What person can do/what they think they can do/what they actually do?
Multi or single item construct?
Item selection
Literature
Patients/experts opinions and experiences
Item banks
Theoretical background: Reflective or formative model (IRT)
Item difficulty (IRT and CTT)
Scoring: Scale/index or profile
Pilot testing
Qualitative methods
Feasibility
Acceptability
Field testing
Item reduction, item difficulty
dimensionality, scale scoring
item functioning
Evaluating psychometric properties of the measure

FIGURE 3 Measure development process

2.3.1 Conceptual framework and construct definition

Construct. Based on the need, the conceptual model or framework should be clearly defined, considering all the aspects of the concept. A conceptual model may be used if the concept or construct can be defined before planning the questionnaires content and possible multiple dimensions. The construct can also be defined post-hoc, using factor analysis, but might not be an efficient way to define it (de Vet et al. 2011, p. 33). In the book of de Vet et al. (2011) reflective and formative model behind the construct and item formulation is introduced. Depending on the construct, either reflective or formative model of the construct has to be chosen to be able to use proper methods for final item selection. As the formative model describes the construct as a result of the items, in the reflective model the items are reflections of the underlying construct (figure 4, modified from de Vet et al. 2011, p. 14).

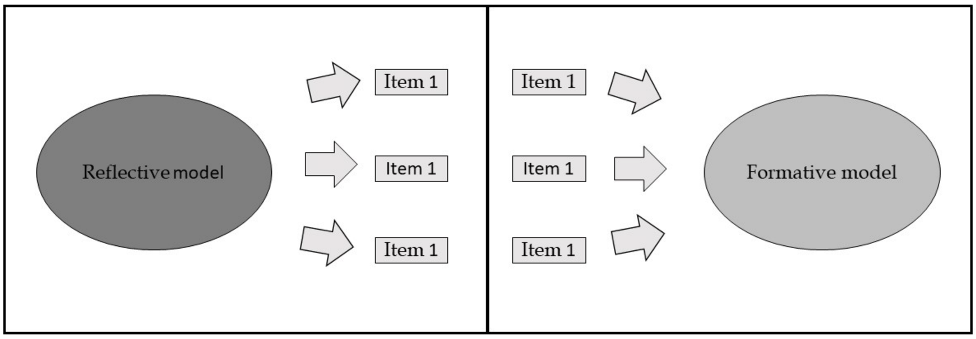


FIGURE 4 Formative and reflective model behind the underlying construct

Target population. The descriptions of the items may vary depending on the target population. Different questions and statements for adults and children are needed. Also, the decision of diseases specific or generic measure has to be made (de Vet et al. 2011, p. 34).

Purpose of measurement. The measures are developed to diagnose, to evaluate and to predict the patient status. For diagnosis, the discriminative properties of the measures are essential at a certain point of time. For evaluative measure, the questionnaire has to have the ability to show the changes over time, and for predictive purposes, the measure is planned to classify or categorize people based on their prognosis. All three properties may also be present in the same measure (Guyatt et al. 1992, de Vet et al. 2011, p. 54-55).

Choice of the measurement method differs in correspondence to the construct. In order to measure the same construct, it is meaningful to separate the objective measurement from the subjective measurement. The information of what a person can do is different compared to what a person thinks they can do and again compared to what they do. The "can-do" measures capacity and the “think they can do” measures the perceived ability while the "they do" measures the performance (WHO 2001, p. 14-15, de Vet et al. 2011, p. 35). Also, the underlying construct defines if single or multi-item questionnaire comes to question (Sloan et al. 2002).

2.3.2 Selecting and formulating items and item reduction

Extensive literature review of similar previous measures helps to choose the items for multi-item questionnaires. An exception is if the measure is developed, e.g. for a purely new disease. According to Cella et al. (2007), it is also possible to use the item banks as Patient Reported Outcome Measure Information System (PROMIS), where the items Item Characteristic Curves (ICC) have been determined using item response theory (IRT). Items for the item banks are also collected from previously developed and used measures (Cella et al. 2007). Using clinicians and patients as experts, who has the experience and knowledge of a specific construct, may be used as informants. Focus group interview is a useful method to collect information from the informants for the new items (Carey et al. 2012, p. 15-18). Several basic rules as unambiguous, positive wording, specific time determination and one question or statement per item has to be considered while formulating the items (de Vet et al. 2011, p. 41).

When the reflective model is behind the defined latent construct, (figure 4) the items are expected to correlate between each other and followed by that, are interchangeable. Thereby the items which are reflecting the same construct may be reduced or replaced with a similar item. To be sure all the items reflecting the underlying construct, are present, the developers need to be sure that enough items are considered to represent the latent construct. Based on the reflective model, the evaluative measure is adopting both discriminative properties between individuals and the ability to detect change over time within the subjects in the same construct (de Vet et al. 2011, p. 13-15). There are conflicting views over these concurrent abilities (discrimination and change detection) related to the validity of the measure. Guyatt et al. (1992) present the concept of signal and noise ratio when exploring the reliability and responsiveness of evaluative and discriminative measure. The item, which is showing good discrimination ability, is not necessarily able to measure the change (Guyatt et al. 1992).

The item difficulty and required response options can be revised using classical test theory (CTT) or IRT. Also, the discrimination ability is studied using IRT based item characteristic curve (Thorpe & Favia 2012).

2.3.3 Item and scale scoring

The chosen measurement level (nominal, ordinal, interval or ratio) either allows or prevents the possibility for a researcher to recode the existing data on other measurement levels. The higher the chosen level is, the more options there are to use a lower measurement level, if needed (de Vet et al. 2011, p. 48-49). In the ordinal scale, the maximum meaningful number of categories in a clinical perspective is found to be seven based on the short and long term memory capability (Miller 1994). de Vet (2011) and Yu (2017) explains the usage of CTT and IRT in scale development. Both CTT and IRT analysis can be used to detect the usage and information of the chosen categories in a specific population. CTT can be used to find out if there are categories which are not selected by any of the respondents. IRT analysis is a proper method to detect the respondents with different abilities in a latent trait, probability of choosing a specific score. Also, the item characteristic curve gives information on the discriminative ability of the item. If the items differ in their discrimination ability, the discrimination parameter is used to weigh the items before summing the scores (de Vet et al. 2011, p. 68-69, Yu 2017).

2.3.4 Pilot testing and field testing - an essential part of the measure development

Hak et al. (2006) and de Vet et al. (2011) are highlighting the importance of the pilot – and field tests of the measure. Measure development needs numerous test phases. Pilot testing relies on qualitative methods and is conducted with a small number of people who are aware of developed construct, but also with the final target group to get the measure understandable and pertinent entirety. Different qualitative methods and their combinations like "The Three-Step Test-Interview" may be used to collect the data to improve the questionnaire. Of importance is to test the feasibility of acceptability of the questionnaire with the target group (Hak et al. 2006, de Vet et al. 2011, p. 58-59). de Vet et al. (2011) goes further on item reduction and obtaining a deep understanding of the data structure. By way of field testing, the dimensions of the construct should be obtained in multi-item measures. To be able to conduct legitimate field tests using quantitative analysis, an adequate number of study participants are needed. If in the pilot phase the number of participants in tests was a few dozens, for field-testing, couple hundred of participants are needed. To explore the dimensionality of the data using factor analysis (FA) or IRT based methods for scale scoring and item functioning, the amount of participants has to be high enough (de Vet et al. 2011, p. 65-66, p. 68-69).

2.3.5 Evaluating the psychometric properties of the published measure

Based on the COSMIN (Consensus based Standards for selection of health status Measurement INstrument) study, Mokkink et al. (2010 a, b) published a taxonomy of psychometric properties built out of three domains: Reliability, validity and responsiveness. Under these domains, seven psychometric property aspects were named. The fourth domain "Interpretability", was not really considered as psychometric property of the measure. Table 2 summarizes the psychometric properties and examples of the statistical methods and characteristics to be used when the patient reported outcome measure (PROM) is evaluated.

The terminology and utilized statistical methods differ in the psychometric evaluation literature. Among the other psychometric properties, the responsiveness and interpretability are off great interest what comes to the use of PROMs to evaluate the treatment usefulness. The COSMIN panel ended in definition of responsiveness as follows: "The ability of the instrument to detect (important) change over time in the construct to be measured". The word important was left out from the definition, because it was more related to the interpretation of the change score (Mokkink et al. 2010a). Beaton et al. (2010) defined the taxonomy for the responsiveness introducing three aspects in it. First the *who* axis, where the focus should be on to whom the measure is being used, was introduced. Second they discussed about the *which* part of the responsiveness. Is the evaluation of the responsiveness done with the data of the scores collected at one point of time or over time? Thirdly, they wrote about *what* kind of responsiveness or change is being analysed. The observed change as statistically significant change, or important change as clinically relevant change?

These two taxonomies and terminologies defining the responsiveness rise an important question of the change. de Vet et al (2010, p. 204) states unambiguously, that responsiveness is always calculated as a change between minimum two measurement points. The measurement has to be done in the population, which is assumed to change in the measured construct towards one direction or another over certain time interval. Instead, Beaton et al. (2001) describes five separate perspectives of the change scores depending on how the change is quantified. Most obvious is the change, which is possible to detect by an instrument based on its range. As an example, the 12-item WHODAS minimum total score is 0 and maximum is 48 when the possible range is 0 – 48 giving a 0 as '*Minimum potentially detectable change*'. The second is '*Minimally detectable change*' (MDC) where the change is defined as true change plus error. The MDC can be quantified using distribution based methods utilizing the standard deviation (SD) and known reliability coefficient of the instrument. de

Vet states that MDC is an important value to interpret the change scores (de Vet & Terwee 2010). Beaton et al. (2001) continues by introducing the *'Observed change'* referring to known efficacy and the *'Estimated change'* referring on external standard as patients or clinicians own evaluation of the improvement. The last perspective of the change, is the *important change.* Here, the question is, who interprets, what is important change? The measure can probably detect change, but is the observed change important to the patient? Or is the experienced or detected change important to the stake holders? Or does the researcher decide, what is the cut-point of important change in hypothese testing, when using the criterion approach?

TABLE 2 Psychometric property domains, aspects, evaluation methods and characteristics

Domain	Psychometric properties	Property aspects	Examples of the statistical methods
Reliability	Internal Consistency		Cronbach's alpha, KR-20, chi square
	Reliability	Test-retest, inter-rater, intra-rater	ICC, kappa or weighted kappa
	Measurement error	Test-retest, inter-rater, intra-rater	SEM, SDC, LoA
Validity	Content validity	Face validity	N/A
	Criterion validity	Concurrent validity, predictive validity	Correlation, ROC
	Construct validity	Structural validity	EFA, CFA, IRT for (uni)dimensionality
		Hypotheses testing	Adequate tests related to hypothese to be tested
		Cross-cultural validity	CTT: CFA, IRT: DIF
Responsiveness	Responsiveness		Adequate tests related to hypothese to be tested (correlation between change scores, mean difference, ROC)
Interpretability			Distribution, mean, SD,

KR-20 = Kuder Richardson, ICC = Intra Class Correlation, SEM = Standard Error of Measurement, SDC = Smallest Detectable Chance, LoA = Limits of Agreement, N/A = Not Applicable, ROC = Receiver Operating Characteristic, EFA = Explorative Factor Analysis, CFA = Confirmatory Factor Analysis, IRT = Item Response Theory, CTT = Classical Test Theory, DIF = Differential Item Functioning, SD = Standard Deviation

2.4 Developing WHODAS 2.0

Construct

The unidimensional conceptual framework of latent trait (functioning) of the WHODAS 2.0 is based on the ICF framework, the ICF Checklist and the ICF Core Sets. Six WHODAS domains cover the dimensions of functioning: a) cognition, b) mobility, c) self-care, d) getting along, e) life activities and f) participating in society (Ustun et al. 2010b, p. 4, Ustun et al. 2010a).

Target population and measurement purposes

The WHODAS 2.0 was developed to evaluate disability and functioning in generic adult population or with individuals in clinical or research use (Ustun et al. 2010b, p. 3-4) for various purposes. It was developed to evaluate the limitations on activities and participation despite the medical diagnosis. Furthermore, it was elaborated to discriminate the disability and functioning levels or to evaluate the effect of health or other interventions (Ustun et al. 2010b, p. 4). The WHODAS 2.0's method to measure functioning is set on the performance level (questioning what person does) and uses multiple items based on six domains linked to an ICF framework (Ustun et al. 2010b, p. 11).

Selecting and formulating the items and item reduction

In order to generate a shared understanding of disability and its variation, focus group and expert interviews were conducted in various countries representing wealthier and poorer societies. The extensive scale review of health assessment methods was utilized to understand the cultural variation in health and disability measure. The continuance towards generic questionnaire included the pooling of the items was based on cross-cultural studies done in 19 countries in Europe, Asia, Africa and North America (Chatterji et al. 2001, p. 21-27 Ustun et al. 2003b, Ustun et al. 2010b, p. 12-13, Ustün et al. 2010a). As a result of Cross-Cultural Applicability (CAR) –study as a pilot study, 96 functioning related items under next six domains were detected: Cognition, mobility, self-care, getting along, life activities and participation. Along with item production, in the CAR–study the WHODAS 2.0 was piloted for cross-cultural validity (Room et al. 1996, Ustun et al. 1999, Chatterji et al. 2001, p. 21-27 Trotter et al. 2001). Several pilot and field studies were conducted to reduce the number of items from 96 to final 36 and the shorter 12-item version of the measure (Ustun et al. 2010 a,b).

Measurement level and Item/scale scoring

In the WHODAS 2.0, each item is evaluated using five response categories in ordinal scale with a verbal description: 1 "None", 2 "Mild", 3 "Moderate", 4 "Severe" and 5 "Extreme or cannot do", or the scoring may be changed in use from 0 to 4. The scoring is completed using a simple or complex method, where simple scoring the given response in each item is summed without weighting of the items. Simple sum score may be reliable if the structure of the tool is unidimensional (Ustun et al. 2010b, p. 19) and has a good internal consistency

(Sloan et al. 2002). Instead, in complex scoring, the IRT based difficulty level of each item is noticed.

Furthermore, the sum score is defined based on weighed scores and the severity of each given response. Complex scoring utilizes a computer program. In addition to the total score, in the 36-item version, the domain scores can be calculated for each domain using a simple or complex scoring method. Based on the pilot studies completed in 19 countries (n=1431), the population norms of a total score of the WHODAS 2.0 are available for both, 36-item and shorter 12-item version (Ustun et al. 2003a, p. 768, Andrews et al. 2009, Ustun et al. 2010b, p. 43-44).

2.5 WHODAS 2.0 pilot and field testing

Followed by an extensive review of disability and health measures, the experts grouped the pooled data of possible items on six existing WHODAS domains: cognition, mobility, self-care, getting along, life activities and participation (Ustun et al. 2010b, p. 12-13). The meaningfulness, universality and cross-cultural feasibility of the pooled and categorized items were investigated in the Cross-cultural applicability (CAR)–study, which was conducted using several qualitative and quantitative methods. Utilizing the results of the CAR-study, the 96 items WHODAS 2.0 version was developed for field testing (Room et al. 1996, Ustun et al. 1999, Chatterji et al. 2001, p. 21-27, Trotter et al. 2001, Ustun et al. 2001, p. 320).

Ustun et al. (2010) describe the two phases of WHODAS field testing: The first phase was aiming at item reduction and feasibility. Phase I studies used the 96-item version of WHODAS 2.0 to find out, which items were dispensable or unnecessary, to investigate, how the shorter version performed and to explore the rating scales feasibility. The study samples varied from 43 to 283 study participants at 21 different sites. The samples consisted of different physical and mental health statuses. The final item selection was based on qualitative field studies, analysis of missing values, item factor loadings, minimal cross-loading, high discriminative ability of the items at the whole range of disability and only slight overlapping of the items (Ustun et al. 2010a, Ustun et al. 2010b, p. 14-16).

The second phase focused on reliability and validity studies. Several psychometric properties were tested at 16 study sites around the world. The study samples in different study sites varied from 57 to 140 consisting of the general population, people with physical, mental or emotional problems and alcohol or substance use. Both genders were recruited for studies from 18 years old or above.

Reliability was tested using test-retest design and results were presented as intra-class correlation coefficient and kappa values. Item total correlation and Cronbach's alpha was used to detect domains and items internal consistency. In these studies Cronbach's alpha varied between 0.79 and 0.98 in six disability domains the total internal consistency in 36-item WHODAS 2.0 being 0.96 (Ustun et al. 2010b, p. 19, Ustun et al. 2010a).

The principal component analysis was utilized to reveal the construct validity of the measure. The analysis confirmed the hierarchical structure of general disability factor and six factors representing the disability domains. All the factor loadings were >0.77 (Andrews et al. 2009, Ustün et al. 2010). Factor structure of the items and domains were also explored in the second phase of field studies using confirmatory factor analysis. The results remained the same as in phase 1 factor analysis (Ustun et al. 2010b, p. 21).

Furthermore, to detect responsiveness and concurrent validity, the 36-item WHODAS 2.0 version was used in Chwastiak and Von Korf (2003) and Ustun's (2010a) reports. The field-testing was conducted amongst various patient populations in nine different countries. The results showed highest correlations between the WHODAS 2.0 total score and measures, which were developed to measure functioning and disability as London Handicap Scale (r=0.75), WHO Quality of Life Measure (r=0.68) and Functional Independent Measure (r=0.68). As anticipated, the correlation between the Short Form Health Survey (SF) mental health part and WHODAS was low, (r=0.17) because SF measures signs of illness, not the disability, solely. Responsiveness to change was evaluated using repeated measures design and calculating effect sizes from the change in mean divided by standard deviation at baseline. The effect size for the WHODAS total score in back pain patients was 0.6 and people with depression 0.65. The difference between baseline and follow-up was statistically significant in both populations (Chwastiak & Von Korff 2003, Ustun et al. 2010a).

For the scoring, dichotomized version (0 denoting no limitations and 1, 2, 3 or 4 denoting any limitation) and the polytomous version where 1 denoting none (limitation), 2 denoting mild, 3 denoting moderate, 4 denoting severe and 5 denoting extreme limitation were tested utilizing Rasch model and polytomous partial credit model (PCM) derived from Rasch model. Both scoring methods were found to be compatible. Also, the IRT based population norms for 36 and 12-item WHODAS 2.0 were derived during the field studies. Getting 22 points out of 100 sets an individual in the 80th percentile of the whole population in 36 –item WHODAS and receiving 11 out of 100 sets an individual in 82nd percentile of the studied population (Ustun et al. 2003a, p. 767-768, Ustun et al. 2010b, p. 22, Ustün et al. 2010a).

2.6 Summary of the newest literature of WHODAS 2.0

As described in the previous chapters, extensive psychometric research was done during the WHODAS 2.0 development phase. Federici et al. (2009b) conducted their first literature review on WHODAS II psychometric properties and updated the review on 2017. Based on Federici et al. (2009b) work, WHODAS has been utilized extensively as a PROM in numerous studies, and its psychometric properties have been studied continuously. However, in their 2009 review, only eight studies published between 2000 and 2008 were reported to be psychometric researches. In these studies, the 6-domain factor structure was confirmed using factor analysis. Based on the internal consistency expressed mostly by Cronbach's alpha, the WHODAS II was found to be reliable, the reliability coefficient varying between satisfactory and high.

Furthermore, the measure was found to be valid based on construct validity, convergent validity and discriminant validity (Yoon et al. 2004, Chisolm et al. 2005, Pösl et al. 2007, Buist-Bouwman et al. 2008, Von Korff et al. 2008). Pösl et al. (2007) also reported the WHODAS II responsiveness as a sensitivity to change. The statistical significance between two measures of the total score was found in all studied diagnose groups except in the stroke population (Pösl et al. 2007).

In Federici et al. (2017) review, 49 of the 810 studies are expressing psychometric properties of the WHODAS 2.0. Of those, 37 were published after Federici's (2009) first review. The studies reported alpha values for a total score between good (0.80 to 0.90) and excellent (>0.90). For 36–item WHODAS, the factor structure was unidimensional and consisted of two-level factor structure with one general disability factor and six second level disability domains (Ustun et al. 2010a). The similar factor structure could not be supported for the post-stroke population (Kucukdeveci et al. 2010). For 12-item WHODAS the one general disability structure was confirmed in several studies (Luciano et al. 2010c, Kirchberger et al. 2014, Carlozzi et al. 2015). The reliability of WHODAS 2.0 was explored using test-retest setting. The intraclass correlation coefficient amongst different populations varied between 0.92 (schizophrenia) and 0.95 (inflammatory arthritis) (Baron et al. 2008, Guilera et al. 2012). The Federici's (2017) review presented dozens of studies reporting concurrent validity between WHODAS and World Health Organization Quality of Life (WHOQL) and WHODAS and SF 36 both cases showing inverse correlations ranging from -0.41 to -0.70 (Federici et al. 2017).

To conclude, only thirteen of 37 studies reporting psychometric properties of WHODAS were studying 12-item version, and of those, one of the studied population consisted of arthritis patients. Three had a general population consisting of different diagnoses or were classified as older people (Federici et al. 2017). None of the studies reported 12-item WHODAS 2.0 psychometric properties amongst musculoskeletal population only, and the critical value, minimal clinically important difference score of the WHODAS 2.0 has not been confirmed.

3 PURPOSE OF THE STUDY

The purpose of this dissertation was to explore the validity and interpretability of the 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain. This study is part of the more extensive validity and implementation of the International Classification of Functioning, Disability and Health (ICF) project.

The next specific research questions will be covered.

1. What is the available evidence of the 12-item WHODAS 2.0 psychometric properties in the general population with non-acute physical disabilities? (Study I)
2. Does a floor or ceiling effect exist in the 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain? (Study II)
3. Is there a significant gender-related differential item functioning (DIF) in the 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain? (Study III)
4. What is the minimal detectable change and minimal clinically important difference in the 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain? (Study IV)

4 RESEARCH METHODOLOGY

4.1 Research process and dissertation structure

This thesis consists of a systematic review (Study I) and three original articles (Studies II-IV). The studies II-IV utilize the cross-sectional data of the larger ongoing project “Validity and implementation of International Classification of Functioning, Disability and Health (Turku ICF-study)”. After collecting the data in Turku ICF-study project, the systematic review (Study I) was conducted to find out the most important missing psychometric properties of the 12-item WHODAS 2.0. Based on the review and the data properties, the research questions II, III and IV were formulated. Figure 5 presents the research design of this thesis and table 3 presents the study designs, population, main diagnosis and psychometric properties under research.

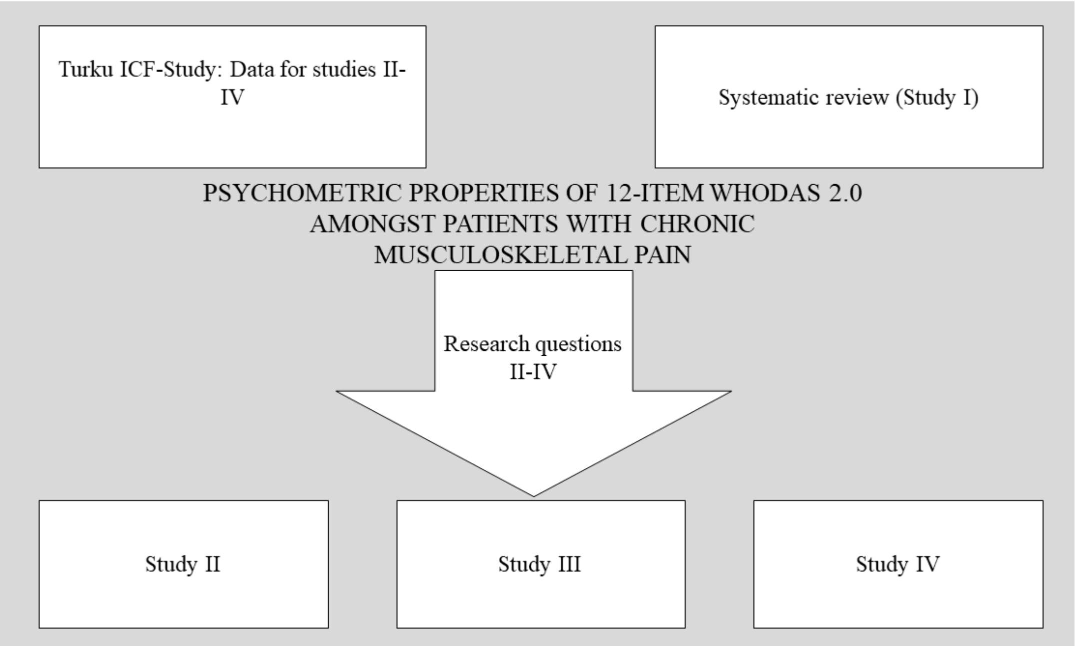

Turku ICF-Study: Data for studies II-IV
Systematic review (Study I)
PSYCHOMETRIC PROPERTIES OF 12-ITEM WHODAS 2.0 AMONGST PATIENTS WITH CHRONIC MUSCULOSKELETAL PAIN
Research questions II-IV
Study II
Study III
Study IV

FIGURE 5 Research process based on the data of the Turku ICF-study and the systematic review

TABLE 3 Description of the study designs, participants, diagnoses and psychometric properties

Study	Design	Participants (n) /number of studies	Main diagnose	Psychometric properties
I	Systematic review	59,408/14	Patients with a non-acute physical disability	Convergent validity, concurrent validity, responsiveness, structural validity, discriminative ability, test-retest reliability
II	Cross-sectional	1,988	Chronic musculoskeletal pain	Interpretability
III	Cross-sectional	1,988	Chronic musculoskeletal pain	Construct validity
IV	Cross-sectional	1,988	Chronic musculoskeletal pain	Interpretability

4.2 Study I Systematic review

Systematic review for this dissertation was conducted following the Preferred Reporting Items for Systematic Reviews (PRISMA) described by Moher (Moher et al. 2009).

4.2.1 Eligibility criteria, literature search and study selection

For the literature search, the original articles, short communications and letters to editors published in peer-reviewed academic journals were searched without language or time of publication restrictions. Conference papers and theses were excluded. The search was allocated in Medline, Embase, Web of Science, Scopus and PsycINFO databases in January 2019. The papers focusing on primary psychiatric diagnoses, acute traumas and other acute conditions, hearing loss and progressive neurological disorders and the population under 19 years of age were excluded. To obtain comprehensive result in literature search, the search clauses were left very generic. An example of search strategy with clause and filters is presented in table 4. The presented search clause was adopted in all five databases. A detailed information of the search process is given in Study I. Title and abstract screening was conducted by two independent reviewer teams and conflicts were solved by consensus or by the referee, the third reviewer. Full-text screening was completed similarly.

TABLE 4 Example of the search strategy and the search clause in Study I

Database	Search clause 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])

4.2.2 Data extraction

For the qualitative analysis of 12-item WHODAS, the following psychometric properties were extracted by one reviewer: validity, reliability, responsiveness and interpretability as minimal detectable change (MDC) and minimal clinically important difference (MCID). The predefined researcher designed form was used to collect the data. The form included information of title of the study, first author, publication year, country of origin, the setting of the study, sample size, gender distribution, age, primary psychometric measure, main quantitative result, main diagnose if possible and the conclusions of the study.

4.3 Construct validity (Study III) and interpretability (Study II and IV) of 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain

4.3.1 Study design and participants (Studies II-IV)

The analysis and results of studies II-IV were derived from the cross-sectional data collected as part of the Turku ICF-project, in the Hospital District of Southwest Finland. Turku ICF study aims to develop the primary clinical practice in Turku University Hospitals (Tyks) physical rehabilitation medicine (PRM) clinic. Hospital District of Southwest Finland consists of 28 member-municipalities. Over 200 000 patients use Tyks's services each year. The PRM clinic is one of the seven outpatient clinics of the operational division of Diseases of Musculoskeletal System. The patients are referred to physical rehabilitation medicine department from primary health care units, occupational health care units and other specialist clinics. The physician's referral is always needed for the visit. The majority of the patients referred to PRM clinic are chronic state musculoskeletal patients; instead, the acute state patients are treated in other clinics. Most of the PRM clinic patients visit there only once. The most common reason for the PRM clinic referral is to get the plan for further examinations or to get the functioning- and work ability evaluation. The majority of the patients return to the primary health care or occupational health care units. Only in single cases the further examination continues in other specialists' clinics.

Of 3,150 patients visiting the PRM clinic between April 2014 and February 2017, 1,988 (63%) participated in the study. Participants were volunteer women and men. The Ethics committee of the Hospital District of Southwest Finland approved the Turku ICF study protocol (ETMK 60/180/2012). Written informed consent was obtained from all participants. The characteristics of the study participants are presented in the chapter 5.2.1 and in the table 5.

4.3.2 Questionnaire

Researcher-designed patient-reported questionnaire was used to record the data of participants. The questionnaire was sent to the patients a few weeks before their clinic visit. It was filled out either at home or in the clinic lobby, before the physician visit. It included a Finnish translation of the 12-item WHODAS 2.0 (Paltamaa, 2014) and questions concerning pain, body weight and height, educational level, perceived general health and age.

The self-administered 12-item WHODAS 2.0

The self-administered 12-item WHODAS 2.0 contained next items covering the most common limitations of functioning appearing in the general population during the last 30 days: 1. standing for long periods, 2. taking care of household responsibilities, 3. learning a new task, 4. joining in community activities, 5. emotional affection by health problems, 6. concentrating doing something for 10 min, 7. walking long-distance as 1 km, 8. washing whole body, 9. getting dressed, 10. dealing with the people you do not know, 11. maintaining friendship and 12. day to day work/school. A Likert-like scale in each item is used to define the severity of the limitation with 0 denoting "no limitation" and 4 denoting "extreme limitation or inability to function". The total score was the sum of all 12 items where a score of 48 points represents the worst possible restriction. The Finnish translation of the 12-item WHODAS is presented in the publication of Paltamaa (2014).

Pain, body mass index, educational level, general health and age

Pain intensity was defined using eleven point's numeric rating scale (NRS) 0 denoting "no pain" and 10 "denoting worst possible pain". Educational level was defined as "further education (which equals in secondary education in Finland)" vs "no further education (which equals no secondary education in Finland)", and body mass index (BMI) was defined as the patient-reported body mass divided by the square of the body height. It was expressed in units of kg/m^2. Perceived general health was defined using five points qualitative scale describing general health as "good", "somewhat good", "average", "somewhat bad" and "bad". Descriptions were quantified as 0 denoting "good" and five denoting "bad". Age was defined in full years at the time of visiting the clinic.

Medical diagnosis
Participant's medical diagnosis was available for researchers from Turku University Hospital's medical record.

4.3.3 Statistical methods

Studies II-IV. The data in basic characteristics were presented as means, standard deviations (SD), medians, interquartile ranges and percentage (%) when appropriate. Statistical comparisons between the groups (genders) in WHODAS total score, pain score, educational level and BMI were made by using t-test.

In publication II and III, the WHODAS total score is presented as percentage (%) of the total score and in publication IV as points. For the clarity, in this conclusion's results section, in study II and III, the total scores are translated from percentage to the points using formula (percentage of WHODAS total score / 100) x 48.

Study II. In case of a rough 5-point Likert-type scale used in WHODAS 2.0 individual items, the ceiling and floor effects of WHODAS 2.0 were calculated numerically as a relative frequency of lowest or highest possible score achieved by the respondents. The cut-off for a significant floor or ceiling effect was set at ≥15%. The distribution of a continuous WHODAS 2.0 total score was analyzed graphically. The probit plotting method was used to detect the non-normality of the WHODAS 2.0 total score's distribution.

Study III. To assess a differential item functioning (DIF) in study III, WHODAS 2.0 items were dichotomized as 'none' (rated by respondents as '0') versus 'any limitation' (rated by respondents as '1', '2', '3', or '4'). The IRT analysis defined discrimination and difficulty parameters of a questionnaire. In this study, discrimination of 0.01 to 0.24 was considered 'none' (a totally level regression curve), 0.25 to 0.64 'low', 0.65 to 1.34 'moderate', 1.35 to 1.69 'high', and discrimination >1.69 was considered 'perfect' (Baker F 2001, p. 34). The probit logistic regression was used to test whether an item exhibits either uniform or non-uniform DIF between gender groups (Swaminathan & Rogers 1990). A two-tailed p-value =<0.05 indicated a significant difference between genders. When significant DIF was observed, the results of the DIF analysis were also presented and evaluated graphically as item characteristic curves based on 2-parameter IRT analysis of dichotomized responses.

Study IV. To describe the variability between an individual's observed score and the true score in study IV, the standard error of the measurement (SEM) was calculated as SEM = SD x $\sqrt{(1 - r_{xx})}$ where r_{xx} is reliability coefficient of the test – in this case, Cronbach's alpha. As the data were cross-sectional, a distribution-based approach was employed to estimate MCID for WHODAS 2.0. Three different formulas were used: 1) MCID = SEM, 2) MCID = 0.5 x SD and 3) MCID = 0.33 x SD. The MDC was calculated as 1.96 x SEM x $\sqrt{2}$. The MDC was also expressed as a percentage (MDC%) – an estimate that is independent of the units of measurement. Representing the relative amount of random measurement error, the MDC % was calculated as (MDC/mean WHODAS total score) x 100. The MDC% <30% was considered acceptable and <10% excellent.

Statistical software. All the analyses were conducted using Stata/IC Statistical Software: Release 15. College Station (StataCorp LP, TX, USA).

5 RESULTS

5.1 A systematic review (Study I)

5.1.1 Study selection

A total of 191 records resulted from the literature search. After removing the duplicates, 86 were screened for eligibility. Of those, 43 studies were excluded for the following reasons: 3 for hearing loss, 20 for psychiatric disorders, 3 for trauma, 2 for Huntington disease, 3 for postpartum or pregnancy and 12 records for 36–item WHODAS and general comments. Based on the title and abstract screening, 13 studies were excluded and remaining 30 were assessed based on the full text. Finally, 14 records were included in the data extraction phase. The flow chart of the study selection is described in figure 6.

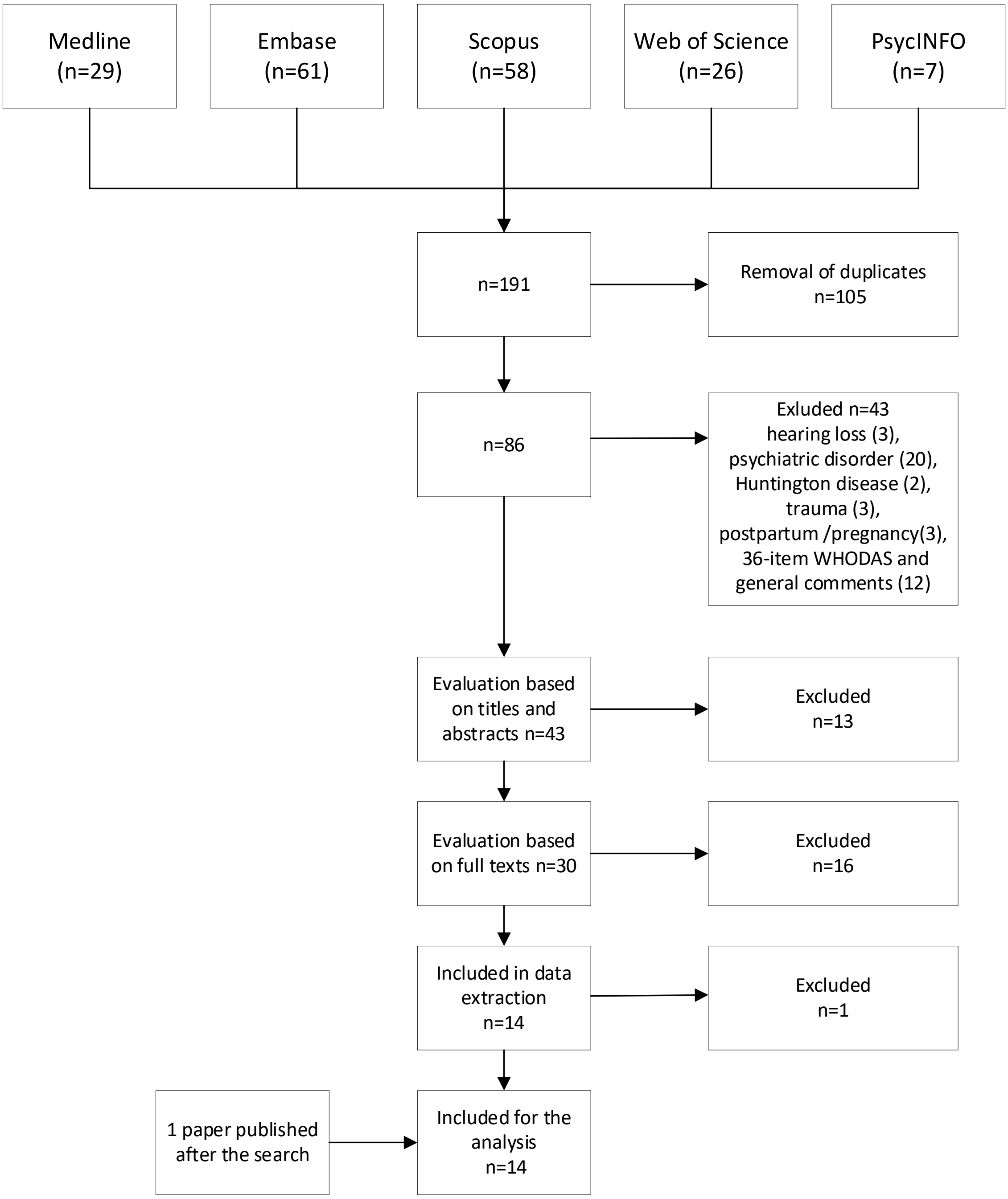

Medline (n=29)
Embase (n=61)
Scopus (n=58)
Web of Science (n=26)
PsycINFO (n=7)
n=191
Removal of duplicates n=105
n=86
Exluded n=43 hearing loss (3), psychiatric disorder (20), Huntington disease (2), trauma (3), postpartum /pregnancy(3), 36-item WHODAS and general comments (12)
Evaluation based on titles and abstracts n=43
Excluded n=13
Evaluation based on full texts n=30
Excluded n=16
Included in data extraction n=14
Excluded n=1
1 paper published after the search
Included for the analysis n=14

FIGURE 6 Flow chart of the study selection in systematic review (Study I)

5.1.2 Population in selected articles

All the 14 papers were observational studies and published after 2013. Of 14 records, 12 papers focused on working-age people, two for the elderly. The sample size varied between 80 to 31 251 study participants, and the majority of the participants were women (from 47% to 98%) The mean age of participants varied between 41.6 (SD 12.4) and 82.3 (SD 8.1) years. A noticeable heterogeneity of health problems was observed among the participants; patients waiting for elective neuro or joint surgery, patients with musculoskeletal pain, fibromyalgia or spinal cord injury, general population, healthy individuals and people compensated for any disabilities.

5.1.3 Psychometric properties of 12-item WHODAS

Reliability of the 12-item WHODAS 2.0 was most commonly reported using Cronbach's alpha. The alpha estimates varied from 0.81 to 0.96. The test-retest reliability was evaluated in two studies. The differences between the repeated measurements were insignificant.

Construct validity of 12-item WHODAS 2.0 scores were most commonly reported presenting convergent validity. Most studies showed high correlations between WHODAS 2.0 and other disability scales, indicating the two scales measuring the same latent trait. In six of the studies, construct validity was evaluated using either exploratory factor analysis or principal component analysis showing a multidimensional structure of the scale. In four of the studies, the confirmatory factor analysis was used. The result varied from poor to good fit of the model. The number of factors varied from one common (disability) factor and six sub-factors to one or two common factors only.

Discriminative validity was assessed using independent samples t-test for Karnofsky Performance Status scale as disability indicator in one study. Another study used item response theory to examine the discrimination of the items. Results were positive and varied from high to perfect.

The distribution of the 12-Item WHODAS 2.0 scores was evaluated in three studies reporting floor and ceiling effect. Two of the studies reported substantial (32%) to significant (80%) floor effect. None of the studies reported a ceiling effect of WHODAS 2.0

The minimal detectable change or minimal clinically important difference was not reported in any of the studies.

5.2 Construct validity (Study III) and interpretability (Studies II and IV) of the 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain

5.2.1 Characteristics of study participants

The participants' demographic, health and functioning characteristics are presented in the table 5. The participants were 48 (SD 6.3) years old, and 1 297 (65%) were women. The average WHODAS 2.0 total score was 13.1 (SD 9.4), and the average intensity of the pain was 6.3 (SD 2.0) points. Most of the patients (n=1 746, 88%) had a main diagnosis 'M' – 'Diseases of the musculoskeletal system and connective tissue' – according to the International Classification of Diseases version 10. The most frequent diagnoses were 'M54 Dorsalgia' (n=781, 39%) and 'M79 other soft tissue disorders' (n=202, 10%). 67% of participants had no high school education, and 33% informed high school education. The average BMI level within study participants was 27.4 (SD 5.7).

TABLE 5 Characteristics of the study participants in studies II-IV

Variable	Total	Men	Women	p-value
N	1,988 (100%)	691 (35%)	1,297 (65%)	
Demographics				
Age, years, mean (SD)	47.6 (6.3)	47.6 (14.7)	47.5 (15.1)	0.927
No high school, n (%)	1,258 (67)	515 (79)	743 (61)	<0.001
High school, n (%)	609 (33)	136 (21)	473 (39)	
Health status				
Body mass index, mean (SD)	27.4 (5.7)	28.2 (5.1)	27.0 (6.0)	<0.001
Pain, points, mean (SD)	6.3 (2.0)	6.2 (2.0)	6.4 (1.9)	0.005
Diagnoses				
Dorsalgia, n (%)	781 (39)			
Other soft tissue disorders, n (%)	202 (10)			
Functioning				
WHODAS total score, mean (SD)	13.1 (9.4)	13.0	13.0	0.843
WHODAS total score median and (inter quartile range)	12 (6 to 19)			

5.2.2 Floor and ceiling effects of 12-item WHODAS 2.0 (Study II)

The distribution of the WHODAS total score and all item scores separately was examined. A significant floor effect was observed in all twelve WHODAS 2.0 items varying from 15% to 79% (table 6). Figure 7 displays the substantial floor effect for a total score as well. No ceiling effect was detected for any of the WHODAS 2.0 items.

TABLE 6 Summary of the floor and ceiling effects of the WHODAS 2.0 per item

WHODAS item	Floor	Ceiling
	%	%
Standing for long periods	29	0
Taking care of household responsibilities	21	0
Learning new task	74	1
Joining in community activities	46	6
Emotional affection by health problems	15	3
Concentrating doing something for 10 min	56	2
Walking long distance as 1 km	37	14
Washing whole body	51	2
Getting dressed	42	1
Dealing with people you don’t know	79	2
Maintaining friendship	62	2
Day-to day work/school	18	0

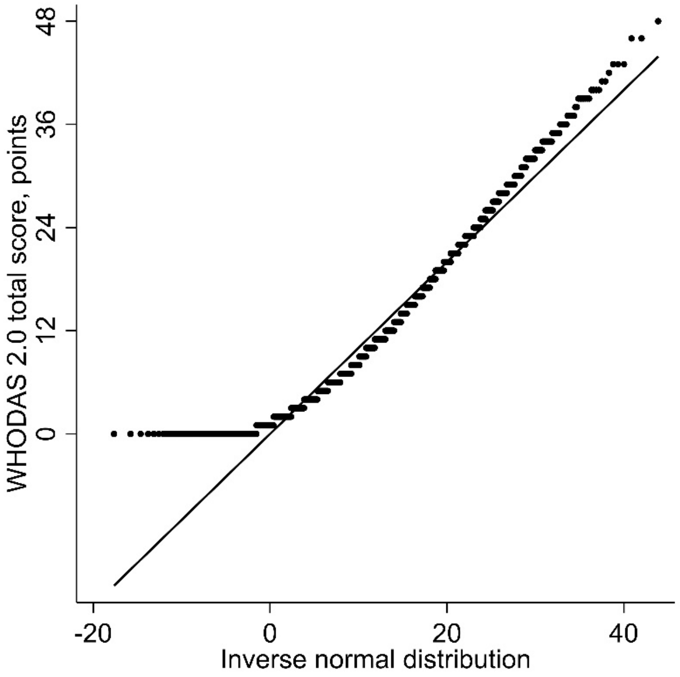


FIGURE 7 Floor and ceiling effects of WHODAS 2.0 total score

5.2.3 Differential item functioning of 12-item WHODAS 2.0 (Study III)

In the discrimination ability of 12-item WHODAS analysis, the total scores of the WHODAS 2.0 were 13.1 (SD 9.4) points (men 13.0 and women 13.1) (p=0.843). The differences between men and women in BMI (p<0.001), pain severity (p=0.005, 95% CI 0.01 to 0.38), and educational level (p<0.001) were statistically significantly different. High to perfect discrimination ability was observed for all the items except for item 9 'dressing' with moderate discrimination (table 7).

TABLE 7 Discrimination values of each WHODAS 2.0 item with 95% CIs

	WHODAS 2.0 Item	Discrimination	95% confidence limits	
			Lower	Upper
1	Standing for long periods	1.45	1.32	1.58
2	Household responsibilities	2.29	2.11	2.48
3	Learning a new task	1.68	1.49	1.87
4	Joining in community activities	2.7	2.47	2.94
5	Emotionally affected by health problems	1.79	1.64	1.94
6	Concentrating	1.92	1.74	2.1
7	Walking a long distance	1.49	1.36	1.63
8	Washing	1.72	1.56	1.88
9	Dressing	1.34	1.21	1.47
10	Dealing with people you don't know	2.05	1.82	2.29
11	Maintaining a friendship	2.18	1.97	2.38
12	Day to day work	1.82	1.67	1.97

Based on the IRT model, the difficulty of the items was studied. The levels of latent ability (functioning) on eight items – '3 learning', 4 'joining in community activities', 6 'concentrating', 7 'walking', 8 'washing', 9 'dressing', 10 'dealing with people you don't know', 11 'maintaining friendship'– were shifted towards the elevated disability level to adopt certain score compared to average disability level of the entire studied population. Other four items 1 'standing', 2 'household responsibilities', 4 'being emotionally affected' and 12 'day to day work' demonstrated a perfect difficulty property (table 8).

In the table 8, the positive values on the ability column are denoting higher disability levels, and the negative values represent the lower disability levels. Interpretation of the results in the table 8 are as follows: If a person's latent ability level in item 1 is -0.87 the score 1 will be endorsed. If the latent ability sets on location -0.19, score 2 will be endorsed. If the latent ability is 0.61, the score 3 will be selected, and the score 4 will be selected, if the latent ability of functioning is located on the level 1.62 on X-axis.

TABLE 8 Difficulty properties of each WHODAS 2.0 item with 95% CIs, the ability column presenting per score the latent ability location on X-axis in IRT analysis item characteristic curve

	1 Standing				2 Household		
Score	Ability	95% CI		Score	Ability	95% CI	
>=1	-0.87	-0.98	-0.76	>=1	-1.02	-1.11	-0.93
>=2	-0.19	-0.28	-0.11	>=2[a]	-0.12	-0.19	-0.05
>=3	0.61	0.52	0.7	>=3	0.78	0.7	0.86
4	1.62	1.47	1.76	4	1.89	1.76	2.03
	3 Learning new				4 Joining community		
	Ability	95% CI			Ability	95% CI	
>=1	0.91	0.81	1.01	>=1	-0.12	-0.19	-0.05
>=2	1.59	1.44	1.74	>=2	0.51	0.44	0.58
>=3	2.23	2.02	2.44	>=3	1.04	0.96	1.12
4	3.29	2.93	3.65	4	1.8	1.68	1.93
	5 Emotional affection				6 Concentrating		
	Ability	95% CI			Ability	95% CI	
>=1	-1.47	-1.59	-1.35	>=1	0.19	0.11	0.26
>=2	-0.17	-0.25	-0.09	>=2	0.95	0.86	1.04
>=3	0.74	0.65	0.83	>=3	1.76	1.62	1.9
4	2.6	2.39	2.81	4	2.82	2.57	3.06
	7 Walking				8 Washing		
	Ability	95% CI			Ability	95% CI	
>=1	-0.49	-0.59	-0.4	>=1 [b]	0.04	-0.03	0.12
>=2	0.2	0.12	0.29	>=2	1.04	0.94	1.14
>=3	0.87	0.76	0.97	>=3	1.83	1.68	1.99
4	1.67	1.52	1.82	4	2.96	2.69	3.23
	9 Dressing				10 Dealing with people you don't know		
	Ability	95% CI			Ability	95% CI	
>=1	-0.32	-0.42	-0.23	>=1	1.02	0.92	1.11
>=2	1.05	0.94	1.17	>=2	1.54	1.41	1.67
>=3	2.1	1.9	2.29	>=3	2.08	1.91	2.26
4	3.93	3.49	4.36	4	2.81	2.55	3.08
	11 Maintain friends				12 Day-to-day work		
	Ability	95% CI			Ability	95% CI	
>=1	0.37	0.3	0.44	>=1	-1.28	-1.4	-1.17
>=2	1.1	1.01	1.2	>=2	-0.37	-0.45	-0.28
>=3	1.71	1.58	1.84	>=3	0.44	0.35	0.52
4	2.75	2.51	2.98	4	1.2	1.09	1.31

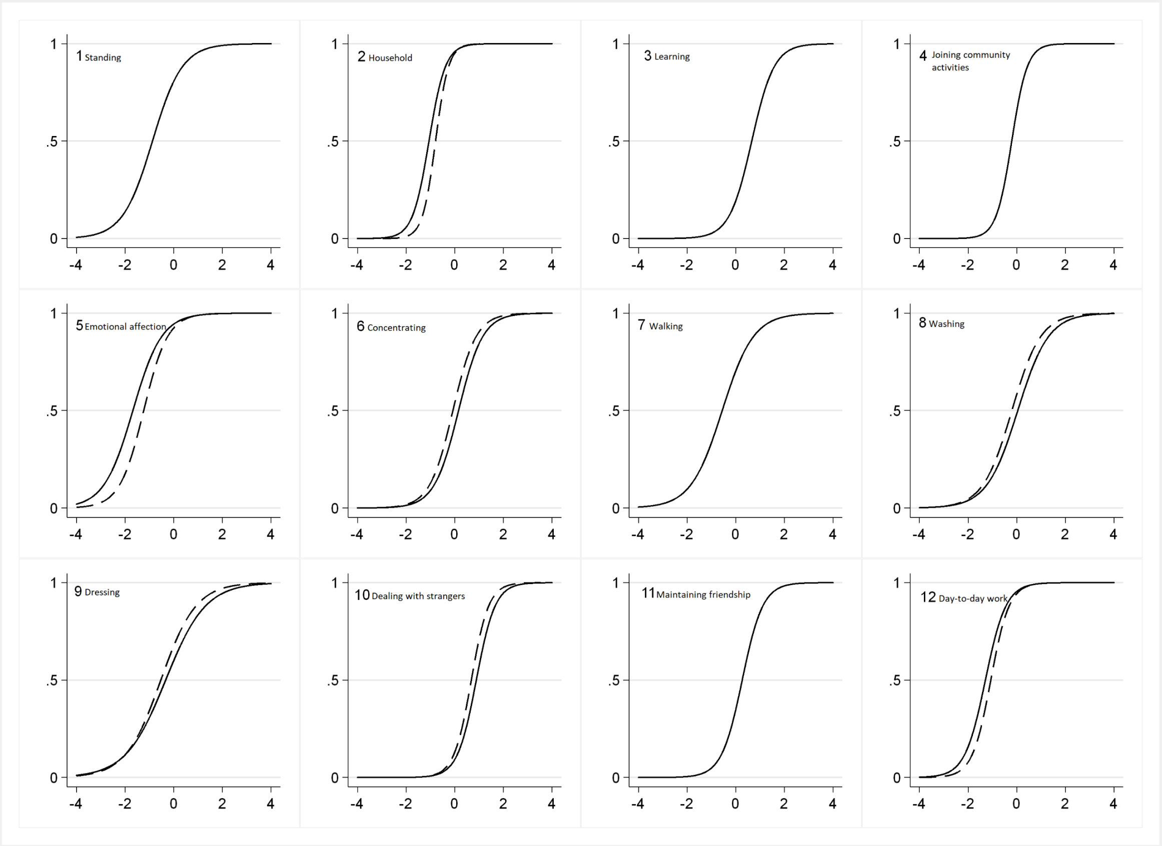
All *p*-values <0.001 except for [a] *p*=0.001 and [b] *p*=0.279.

Significant DIF between the genders was observed in seven out of 12 items: 'household responsibilities', 'being emotionally affected', 'concentrating for 10 minutes', 'washing', 'dressing', 'dealing with people you don't know', and 'work'. All the detected DIFs were uniform (table 9).

TABLE 9 The Chi^2 test results on <0.05 significance level showing WHODAS 2.0 items with uniform differential item functioning between men and women

WHODAS 2.0 items	Chi^2	*p*-value
2 Household responsibilities	10.84	0.001
5 Emotional affection	10.9	0.001
6 Concentrating	13.5	0.0002
8 Washing	12.27	0.0005
9 Dressing	7.57	0.0059
10 Dealing with people you don't know	8.49	0.0036
12 Day-to-day Work	3.97	0.0464

For items 2 'household responsibilities', 5 'emotional affection' and 12 'work', men had to experience slightly worse disability than women to achieve the same score. A reverse effect was observed for items 6, 8, 9 and 10 ('concentrating', 'washing', 'dressing', 'dealing with people you don't know') (figure 8).



Y-axis presents the probability of endorsing the item. X-axis presents the ability level. Estimates for men in dash lines, estimates for women (or the entire sample) in solid lines.

FIGURE 8 Item characteristic curves of WHODAS 2.0 items 1-12 visualizing DIF between the genders and difficulty and discrimination abilities of the items.

5.2.4 Minimal clinically important difference and minimal detectable change of the 12-item WHODAS 2.0 (Study IV)

The distribution of the WHODAS 2.0 total score was abnormal with a shift towards the mild disability levels (figure 9). As the median and mean points were alike, the distribution was considered close to normal enough to proceed with calculations based on mean and standard deviation. The Cronbach's alpha as a measure of internal consistency was good 0.89. The mean WHODAS 2.0 total score was 13.1 (SD 9.4) and median 12 (Interquartile range from 6 to 19, range 0 to 48) points.

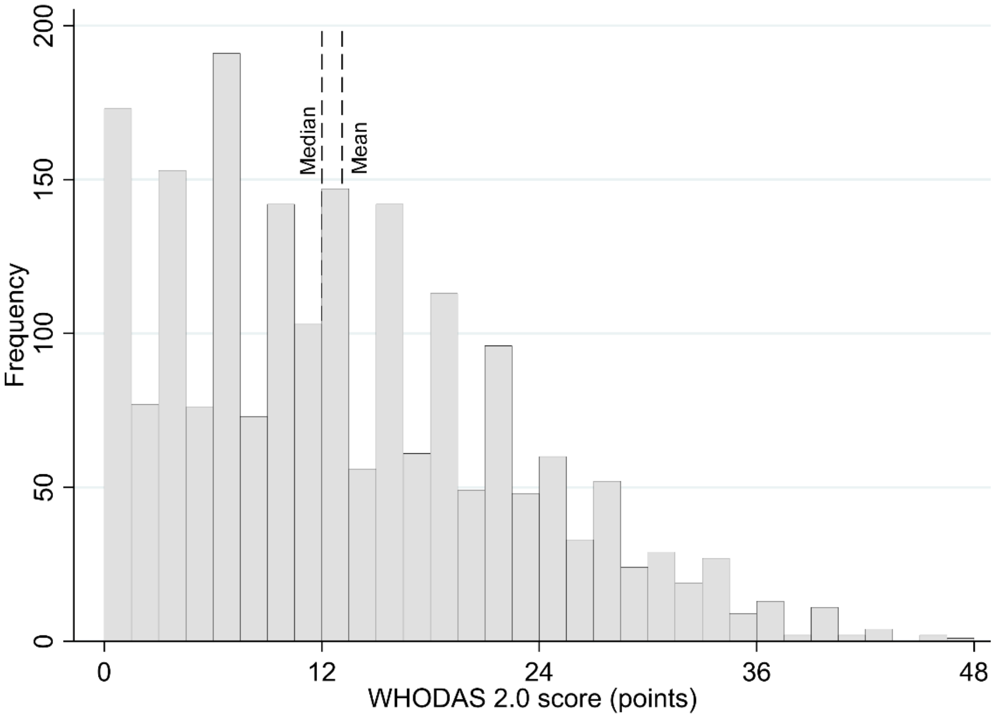


FIGURE 9 The distribution of the WHODAS 2.0 total score amongst people with chronic musculoskeletal pain

The MCID estimates for WHODAS 2.0 ranged from 3.1 (calculated as SEM as 1/3 of SD) to 4.7 points, respectively. The MDC was 8.6 points, and MDC% was 66%.

6 DISCUSSION

The main findings of this thesis were that amongst the chronic musculoskeletal pain patients, the 12-item WHODAS 2.0 demonstrated a high MDC of almost nine points. As the MDC exceeded the level of MCID, nine points were considered to be the amount of change perceived by a respondent as clinically significant. A significant floor effect (>15%) was seen in all 12 items. A substantial floor effect for a total score was also detected graphically. A significant uniform gender-related DIF was detected in seven out of 12 items.

6.1 Methodological considerations

The strength of the study was the large study sample, almost 2000 chronic musculoskeletal patients who completed the 12-item WHODAS 2.0 questionnaire. In the systematic review, the number of participants varied between 80 to 31 251 participants (mean n 3 969, median 183). Furthermore, compared to the studies conducted during the development of the WHODAS 2.0 (Ustün et al. 2010 a,b), in the present studies II-IV the sample was more homogenous and gave a reasonable basis on the specific psychometric evaluation of the measure for the adults with musculoskeletal pain. On the other hand, our sample did not include people with osteoarthritis and rheumatoid arthritis, two of the most common disabling musculoskeletal conditions in the world (Briggs et al. 2018). Instead, Baron et al. (2018) and Kutlay et al. (2011) examines the validity and reliability of WHODAS 2.0 amongst osteoarthritis and rheumatoid arthritis patients, but the 36-item version of the WHODAS 2.0 was explored in both studies (Baron et al. 2008, Kutlay et al. 2011). Thus, according to the large study sample, the results of the present study may be generalized in a population having the main diagnosis of 'Diseases of the musculoskeletal system and connective tissue' and under that, 'Dorsalgia' and 'Other soft tissue disorders'.

From a methodological point of view, it has to be acknowledged that despite the large sample and sampling method, consecutive physical medicine rehabilitation policlinic patients, one fourth reported only 6 points out of 48 WHODAS total score, and three fourths less than 20. It seems that the studied population was not very heterogeneous in regards to measured disability with 12-item WHODAS 2.0. Gaskin et al. (2017) noticed in specific psychometric evaluations that a skewed sample distribution with a high number of participants evaluating their disability level very low resulted in censored data (Gaskin et al. 2017). That might also be the case in regards to results obtained in the study II - Floor and ceiling effects of 12-item WHODAS 2.0. However, the large sample size allows the use of powerful item response theory, which was used in study II to detect the DIF of the 12-item WHODAS 2.0. The two-parameter model requires at least 500 participants for the measure development (Kean & Reilly 2015, p., 197).

To our knowledge, this is the first study to explore interpretability or responsiveness (Beaton et al. 2001, Mokkink et al. 2010a) of the 12-item WHODAS 2.0 score in people with chronic musculoskeletal pain. It is also notable that the recent systematic review of Federici et al. (2017) identified only one study presenting the minimal clinical important difference and minimal detectable change of the WHODAS 2.0 total score for institutionalized ambulatory older adults (Silva et al. 2019). Federici et al. (2017) results are in line with the results of the Study I. The information of the estimates of the minimal detectable change and the minimal clinically important difference are essential when the clinicians are evaluating the usefulness of selected measures and finally, the rehabilitation results.

However, there are certain limitations according to the statistical method used in the current study to explore the MCID and MDC. Part of the literature supports the use of an anchor-based method to detect the minimal clinically important difference in the scale (de Vet & Terwee 2010). The anchor (patient or clinicians' own expression of the change in the functional status, pain or other) is used to detect the correspondence of the anchor and the studied measure. Regardless, the distribution-based methods are also used in psychometric studies (Kohn et al. 2014). The current study utilized distribution-based method, as there was no anchor available as external criteria. Therefore, the interpretation of minimal clinically important difference score on an individual level based on the present study results should be cautious.

In Beaton et al. (2001) extensive non-systematic literature review, the authors have created a tri-axial taxonomy of responsiveness. The taxonomy highlights the multifaceted phenomenon of responsiveness. Beaton et al. (2001) names three axes in the taxonomy: *The first axis, 'Who'* describes to whom the interpretation of the results is utilized, individual or group level. T*he second axis 'Which'*, describes, at which time point the data for the MCID or MDC estimation were gathered; At a single cross-sectional point, when the distribution between persons is utilized, or over a specific time to assess the within-person change. These two concepts of the time point represent different aspects of responsiveness and should not be treated as the same (Beaton et al. 2001). Still, there are studies where the MCID for longitudinal use has been estimated utilizing cross-sectional between-person settings. The results differ from anchor-based methods but may be used to give an estimation of the MCID for within-patient setting (Redelmeier & Lorig 1993, Redelmeier et al. 1996). *Thirdly, the taxonomy refers to the 'What' axis*, which defines various types of changes. As in present study, one of the changes is minimal detectable change, expressing the amount of change in the total score, which is found to be true without the measurement error. In current study, due to the cross-sectional setting, the widely used internal consistency unit, Cronbach's alpha, was utilized instead of the reliability coefficient to calculate standard error of measurement for the minimal detectable change (Haley & Fragala-Pinkham 2006).

However, the present study gives important novel information of the 12-item WHODAS 2.0 psychometric properties, especially the minimal detectable change and differential item functioning, which both are important factors while developing, choosing and interpreting a measure and its results in clinical use amongst patients with musculoskeletal pain.

6.2 A systematic review of the 12-item WHODAS 2.0

To conduct the systematic review of the psychometric properties of the measure, certain guidelines or steps have been proposed. de Vet et al. (2011) lists ten general steps for a systematic review of measurement properties. The list does not differ considerably on the phases of other type of reviews (de Vet et al. 2011, p. 276). Only one systematic review and one non-systematic review of the WHODAS 2.0 have been published before the study I. Federici et al. (2009, 2017) conducted their reviews aiming to answer the general research questions concerning the usage, versions, language, psychometric properties etc. of the WHODAS II and WHODAS 2.0. The psychometric properties of the measure was one question among the others. The research question was not precisely focusing on certain psychometric property of certain version of the measure on certain patient population. Therefore the comparison of Federicis et al (2017) reviews results on current study is limited. In the current study, the focus was primary on the all psychometric properties of the 12-item WHODAS 2.0 and furthermore, the study focused on people with non-acute physical reasons of disabilities. The convergent result with Federici et al. (2017) and current study was that the MDC or MCID of the score were not known. Instead, contradicting results concerning the unidimensionality of the 12-item WHODAS 2.0 structure was found. In Federici et al. (2017) review they found six studies reporting one-factor solution, instead in the current review only one of the nine studies using either EFA, PCA or FA, reported good model fit for unidimensional structure of the 12-item WHODAS 2.0. The methodological limitation in current study was the fact that the methodological evaluation of the included studies was not conducted. Notable is that only four publications explored the psychometric properties of WHODAS within the musculoskeletal patients and three of them utilized the same population. To get wider understanding of the psychometric properties of the 12-item WHODAS in musculoskeletal population, more research should be done.

6.3 Floor and ceiling effects of the 12-item WHODAS 2.0

Generally, the interpretation of the results between different studies was complicated because of the heterogeneous use of the concepts. The linkage of the items and WHODAS 2.0 domains clarifies the comparison of the results (table 10). Part of the studies published before the present study expressed the results using the domain level, and others used the item level. The 12-item WHODAS 2.0 version is comprised of two items (questions) from each six domains of the WHODAS 2.0.

Concerning the second research question, the significant floor effect for 12-item WHODAS 2.0 total score and all twelve different items were found in present study. The highest floor effects were detected under 'cognition' and 'getting along' domains (table 10). As far as we know, this was the first study detecting the floor and ceiling effect of 12-item WHODAS 2.0 total score amongst people with musculoskeletal pain. Therefore, the comparisons on previous studies were made considering the different study population and different versions of the measure. There were only a few reports where the same 12-item measure was studied in the sense of floor and ceiling effect. According to Schneider et al. (2015), their study was conducted amongst maternal depression population. In that study, high, 74% adoption of the lowest quarter of the total score was reported. The results could not be compared as the population differed from current on age, diagnose and gender (Schneider et al. 2015). Again, remarkable floor effect was found in the Kirchberger et al. (2014) register study of myocardial infarction patients whose infarction had been taken place 6.5 (2.4) years ago. The floor effect in all twelve items varied between 28 to 84% depending on the item (Kirchberger et al. 2014). Further on, floor effects between 25 to 88% were detected using web-based 12-item WHODAS amongst people with anxiety and stress disorders (Axelsson et al. 2017).

Similar to the results of the present study, were found in the work of Meesters et al. (2010) amongst older adults diagnosed with rheumatoid arthritis. Without producing an exact percentage, the authors reported significant floor effect under domains 'cognition' and 'getting along' in 36–item WHODAS II. Notable is the small (N=85) study sample size (Meesters et al. 2010), where few high and low-end scorings affect more on changes in sample distribution than in the larger sample size.

Furthermore, significant, 59% floor effect was found in non-disabled people and 39% in various disabled population in 'mobility' domain (in WHODAS II version domain is named 'getting around') of 36–item WHODAS. According to Federici et al. (2009), 74 % of non-disabled and 52% of disabled reported zero points in 'self-care' domain (Federici et al. 2009). In contrary to present Study II, Wolf et al. (2012) and van der Zee et al. (2014) detected > 15%

ceiling effect in the items 'understanding’ and 'communication' (34-53%) and domain 'self-care' (18-28%). The studied population, people with spinal cord injury, completed the 36-item version of WHODAS II (Wolf et al. 2012, van der Zee et al. 2014).

Due to different versions of the WHODAS and heterogeneous population, convergent findings between the previous studies and the current study is difficult to draw. Still, the typical finding is that the scale shows the remarkable floor effect on all 12 items under all six domains of the measure. In present study, the population was not severely disabled. The total score was only 13 out of 48 points. The 12-item WHODAS total score in Study II represented approximately the same total score, which represents 85% of the population's total score (Ustun et al. 2010a). Because the present population was gathered from the physical rehabilitation medicine clinic, where the patients are referred from the primary health care centres to get the evaluation of their working ability, the population showed good functioning and low disability level. At the same time, the findings of the significant floor effect raised the question: Should the scale be more detailed to detect minor differences in the lower end of the scale, or should it be used amongst people with more severe disabilities? In recent study amongst people with low back pain, Cwirlej-Sozanska et al. (2020) did not detect significant floor effect in 36-item WHODAS 2.0 total score. Instead, convergent results with Study II in domain level were found. In Cwirlej-Sozanska et al. study the ‘cognition’ and ‘getting along’ domains exceeded the set 15% floor effect level (Ćwirlej-Sozańska et al. 2020). Modern work highlights the need for learning new tasks and personal capacity for networking with other people. If the WHODAS 2.0 is used to evaluate work ability after the rehabilitation, there might be a risk that the measure is not able to detect the patient improvement, if the scores given at the baseline are already very low when more sensitive answer options are not available. On the other hand, the scale may show better functioning ability than a person has, just because of the floor effect of the measure.

TABLE 10 Linking WHODAS 2.0 domains and items

Domain	#	Item description
Cognition	6	Concentrating on doing something
	3	Learning a new task
Mobility	1	Standing a long period
	7	Walking for long distance (1km)
Self-care	8	Washing your whole body
	9	Getting dressed
Getting along	10	Dealing with people you don't know
	11	Maintaining friendship
Life activities	2	Take care of household responsibilities
	12	Day-to day work or school
Participation	4	Joining community activities
	5	Emotionally affected by your health condition

6.4 Differential item functioning

In order to determine, how the WHODAS items performed in the chronic musculoskeletal patient population, the IRT analysis showed in current study, that in all eleven items of total 12, the discrimination ability was high or perfect and one of the items ('dressing') showed moderate discrimination ability. Further, item difficulty indexes of eight items – (learning, joining in community activities, concentrating, walking, washing, dressing, dealing with the people you don't know and maintaining friendships) – were shifted towards the elevated disability level compared to average disability level of the studied population. The rest four items (standing, household responsibilities, being emotionally affected, and work) demonstrated perfect difficulty properties while the difficulty indexes were evenly distributed around the zero. In the development phase of the scale the items with low discrimination ability, could be considered to be deleted, as the patients with different overall disability may choose the same score for these items. In the current study, the shifted item location towards elevated disability in eight out of 12 items may reveal on a question, if the measure is able to discriminate the patients with all levels of disability, or just between the patients with mild or severe disability.

Significant differential item functioning by genders was found in next seven items: 'household responsibilities', 'being emotionally affected', 'concentrating for 10 minutes', 'washing', 'dressing', 'dealing with people you don't know', and 'work'. All the detected DIFs were uniform. For items 'household responsibilities', 'being emotionally affected' and 'work', men had to experience slightly worse overall disability than women to endorse the same score as women in a given item. An opposite result was found for items 'concentrating', 'washing', 'dressing' and 'dealing with people you don't know'.

These findings identified several possible problems of the 12-item WHODAS 2.0 amongst patients with chronic musculoskeletal pain. If the items are used for screening purposes, the difficulty index of six items under cognition, self-care and getting along -domains shows that the scales are not able to distinguish between the people who have limitations in their functioning from the people who have not. Further on, the items are functioning differently between the genders under life activities, self-care, participation, cognition and getting along –domains. Men have to experience more limitations in their overall functioning to give the same score in the item than women in the household, work and emotional affection –items. The opposite was seen in ʻconcentrationʼ, ʻwashingʼ, ʻdressingʼ and ʻdealing with strangersʼ –items, where women had to experience higher limitations to give the same score as men.

The differential item functioning of 12-item WHODAS 2.0 had been studied earlier in two studies (Luciano et al. 2010c, Kirchberger et al. 2014) without a sign of DIF by gender. Luciano et al. (2010 a,b,c) and Kirchberger et al. (2014) studies represented the population with diagnose of first major depression or myocardial infarction. Kutlay (2011) with her colleagues, studied the differential item functioning amongst knee osteoarthritis patients using 36-item WHODAS. They detected DIF by gender in the item 'getting your household activities done as quickly as needed', which goes under the domain Life activities, similar findings as in Study III showing the DIF in 'taking care of household responsibilities' (Kutlay et al. 2011). Furthermore, the study of Novak et al. (2010) explored the DIF in 10 items in the 16-item WHO-DAS version during the measure development phase. Their results are not comparable to present study, as there existed items, which are not included in 12 or 36-item WHODAS 2.0 measure or could not be placed under any of the six present WHODAS domains.

The results of the current study existing DIF in 12-item WHODAS by gender were contradicting with previous studies of Luciano et al. (2010a) and Kirchberger et al.(2014). The contrary results may be explained by the differences in the studied population, the age group and used statistical method. In Luciano et al (2010c) study, they used kernel-smoothing technique to detect DIF while in Study III, the logistic regression was utilized. DIF can be estimated using several techniques. Moses et al. (2010) compared the most commonly used statistical methods to show DIF; raw data, logistic regression, log-linear models and kernel smoothing. In their study, the logistic regression was found to be most accurate and kernel smoothing technique being least accurate, especially, if the groups differed in their overall ability, which was also the case in Luciano et al. (2010 c) research.

Furthermore, the nature of the population according to the disease was different and maybe affected examinees answers. Myocardial infarction is an acute, life-threatening stage, whereas the chronic musculoskeletal pain may be considered as less vulnerable status. That might have affected women and men experiencing certain items differently in the chronic stage, compared to the situation of myocardial infarction which affects very fundamental way on a person's existence. Therefore, the gender might not be a distinct factor. The theoretical question of DIF existing or not existing in the same item was further discussed in the publication of Zumbo (2007) He raised an approach of "The third generation of DIF" where the test situations affect latent ability. In Luciano et al. (2010 a,b,c) study, the latent ability, the medical state of the participants, should be noticed (Zumbo & Gelin 2005, Zumbo 2007). A recent study of Gomez-Benito et al. (2017) conducted in lower and middle-income countries verged slightly on this approach. They aimed to explore the factors explaining the order of self-rated severity of disability in WHODAS using anchoring vignette questions. Still, their results did not explain gender-related DIF. Gomez-Olive et al. (2017) were studying the overall variation in disability for age and gender using mixed methods as Gomez-Benito et al. (2018) suggests for the use of DIF in validation studies, but they did not detect DIF (Gomez-Olive et al. 2017, Gomez-Benito et al. 2018).

6.5 Minimal detectable change and minimal clinically important difference

The fourth question in this research was to find out the MCID and MDC of the 12-item WHODAS 2.0 to be able to interpret the results of the total score in the future. Amongst patients with chronic musculoskeletal pain, the minimal clinically important difference ranged from 3.1 to 4.7 points, depending on the used formula. The minimal detectable change was 8.6 points equivalent of 66%. To express the importance of these findings, figure 10 visually shows the relationship between MCID and MDC. The illustration in figure 10 is modified from the picture of de Vet & Terwee (2011, p. 260). The full line expresses the possible scores of 12-item WHODAS, which may vary between 0 to 48 points. In the studied population, the overall WHODAS score in the population was 13.1. Considering the situation where a person was evaluated before rehabilitation intervention and after, the change in his WHODAS score should decrease from 13 to 4, or increase from 13 to 22, to show the change to be a real and meaningful change in functioning without a measurement error. From rehabilitee's point, the decrease in the score should be between 8 to 10 to be clinically meaningful, but still there exists the possibility of approximately 4 points of measurement error.

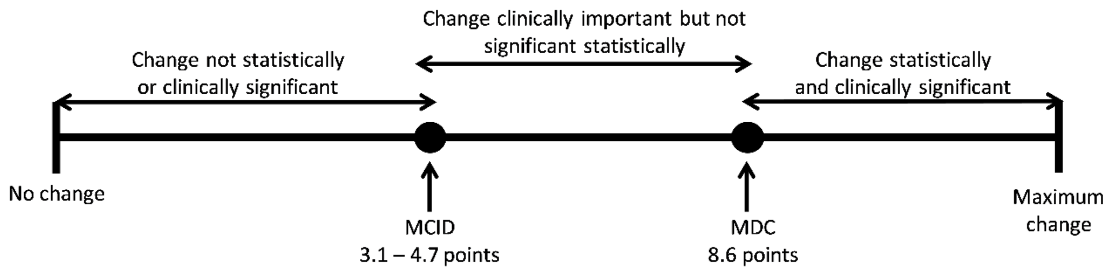

Change clinically important but not significant statistically
Change not statistically or clinically significant
Change statistically and clinically significant
No change
MCID
3.1 – 4.7 points
MDC
8.6 points
Maximum change

FIGURE 10 Statistical and clinical expression of MCID and MDC in 12-item WHODAS 2.0 of chronic musculoskeletal pain patients

For MCID and MDC the present study is as far as we know, one of the few studies where the minimal detectable change and minimal clinically important difference of 12-item WHODAS have been explored. This current study supports the findings of Silva et al. (2019), who reported the 8.15 $MDC_{95\%}$ within elderly participants from nursing home and day-care centre. It has to be acknowledged that in their report, the WHODAS total scores were presented per domain, but the presented sum score did not fit in the expressed domain scores (Silva et al. 2019). In Axelsson et al. (2017) study amongst patients with anxiety and stress disorders, the minimal clinically important difference was close to the results of present study, between 3 and 7. To determine the MCID, Axelsson et al. (2017) used the diagnose specific measure as an anchor.

It is essential to bear in mind that the findings of current study should be cautiously extrapolated to other patients than chronic musculoskeletal pain out-patients. The statistical method used in this present study was based on standard deviation and therefore is sample dependent.

6.6 Practical implications

The findings of this thesis have important implications to understand the interpretation of the 12-item WHODAS scores when deciding on its use in health care and rehabilitation practise. Chronic musculoskeletal pain is a widespread problem but based on the scores of the 12-item WHODAS, the disability seemed to be only mild in the studied population. It should be emphasised that followed by the results of the current study, the high floor effect and difficulty level shifted towards mild disability in several items, the 12-item WHODAS 2.0 can not separate the disability levels in chronic musculoskeletal pain population. Still, musculoskeletal pain is one of the most commonly faced problem in PRM clinic and also in physiotherapy practise. It causes most of the sickness absence in Finland. At the same time, musculoskeletal pain is probably the leading cause of work presentism long before sickness absence. If the 12-item WHODAS is utilized for example on work ability evaluation to observe the functional limitations, the widely detected floor effect of the measure prevents to identify those people, who are actually in the risk of more long-lasting musculoskeletal problems or the decrease in their functioning.

An interesting finding was also the differential item functioning of the 12-item WHODAS. In clinical practice, it underestimates the results of the disabilities in part of the scale in both genders. As noticed before, the reason for the DIF should be explored more in detail to be able to decide if some of the questions should be scaled differently for men and women, or should be dropped out. It is possible that in the studied age group male are slightly more work-oriented and experience social pressure not to "complain" about their work ability even it was seen that the overall disability was similar in men and women. Further, because the questions of the measure are not working similarly between the different subgroups, there should also be different reference values for these groups. For now, the 12-item WHODAS scores are interpreted similarly across the genders. This is not a case, for example, in the measures which are evaluating the objective physical abilities.

Despite of the statistical method estimating the MDC and MCID of the 12-item WHODAS in this study, getting the information of these thresholds is an essential result of this study. From a clinical perspective in individual and societal side, interpreting the change scores is one way of making decisions in work ability evaluation and following the rehabilitation results and making the changes in rehabilitation contents if needed. Knowing the MDC of WHODAS gives an important reference value for this use.

7 MAIN FINDINGS AND CONCLUSIONS

The findings of this dissertation can be summarized as follows:

1. Based on the systematic review, the 12-item WHODAS 2.0 scale showed multidimensionality presenting one common disability factor and several sub factors
2. The 12-item WHODAS 2.0 showed significant, over 15 percentage floor effect in all twelve items. The ceiling effect was not detected in any of the items
3. The 12-item WHODAS 2.0 showed uniform differential item functioning between the genders in seven of twelve items
4. The minimal clinically important difference of the 12-item WHODAS 2.0 was set between 3.1 to 4.7 points and the minimal detectable change was 8.6 points

In conclusion, due to the floor effect, the 12-item WHODAS may have discrimination limitations in the lower end of the scale amongst chronic musculoskeletal patients with milder disabilities. Men had to experience more functional limitations compared to women to achieve the same score for the items defining restrictions in household activities, emotional affection and work. The opposite result was detected for the items defining concentration, washing, dressing and dealing with strangers. High MDC of almost nine points might complicate the use and the interpretation of the total score of WHODAS 2.0 in the work ability and rehabilitation evaluations amongst chronic musculoskeletal pain patients. The 12-item self-administered WHODAS 2.0 appeared to be a multidimensional scale, and its total score may represent different combinations of several contributing factors. Therefore, utilizing the 12-item WHODAS in work ability evaluations or other treatment efficacy evaluations amongst patients with chronic musculoskeletal pain, all the previous findings should be taken into consideration while interpreting the results based on the single total score instead of using item scores. The further research should be conducted to explore the reliability and validity of domain scores in the 12-item version to be able to create a functional profile using the short version of WHODAS 2.0.

YHTEENVETO (FINNISH SUMMARY)

12-osaisen WHODAS 2.0 -mittarin psykometriset ominaisuudet kroonisilla tuki– ja liikuntaelinkipupotilailla

Toimintakyvyn arviointiin on kehitetty luokittelua, malleja ja arviointimenetelmiä vuosikymmenien ajan. Diagnoosilähtöisestä, sairauskeskeisestä arvioinnista on siirrytty toimintakyvyn, toimintarajoitteiden ja terveyden kansainväliseen luokitteluun (ICF), jonka pohjalta Maailman terveysjärjestö WHO julkaisi WHODAS 2.0 Terveyden ja toimintarajoitteiden arviointimenetelmän tämän hetkisen version vuonna 2010. Tätä WHODAS 2.0 12 kysymyksen mittariversiota on suositeltu käytettäväksi aikuisten toimintakyvyn itsearviointiin kuntoutustarpeen tunnistamisessa ja kuntoutumisen seurannassa keväästä 2020 lähtien. Jotta mittarin antamia tuloksia voidaan tulkita luotettavasti, tulee sen mittausominaisuuksista olla riittävästi tietoa. WHODAS 2.0 -mittarista on olemassa useita eri versioita mittarin käyttötavan ja osioiden lukumäärän vaihdellessa. Mittarin pitkää 36 kysymyksen versiota on tutkittu melko kattavasti, mutta lyhyttä 12 kysymyksen versiota vähemmän. Esimerkiksi tietoa mittarin antaman pistemäärän muutoksen suuruudesta, jotta muutos voidaan tulkita todelliseksi muutokseksi toimintakyvyssä mittavirheen sijaan, ei ollut olemassa. Tämän väitöskirjatutkimuksen tarkoituksena oli selvittää 12 kysymyksen itse täytettävän WHODAS 2.0 -mittarin suomenkielisen käännösversion mittariominaisuuksia kroonisilla tuki- ja liikuntaelinkipupotilailla.

Väitöskirjan ensimmäisessä osajulkaisussa tarkasteltiin systemaattisen kirjallisuuskatsauksen avulla 12 kysymyksen WHODAS 2.0 –mittarin psykometrisiä ominaisuuksia henkilöillä, joilla oli kroonisia fyysisiä toimintarajoitteita. Katsauksessa todettiin mittarin olevan luotettava ja korreloivan hyvin muiden toimintakykyä tai toimintarajoitteita arvioivien mittareiden kanssa. Mittarin faktorirakenne todettiin monidimensionaaliseksi.

Väitöskirjatutkimuksen kolmessa seuraavassa osajulkaisussa käytettiin vuosina 2014-2017 fysiatrian poliklinikalla kerättyä poikkileikkausaineistoa (N= 1 988). Aineisto kerättiin käyttäen 12 kysymyksen WHODAS 2.0 -mittaria sekä demografisia tekijöitä selvittävää kyselyä. Kyselyihin vastanneiden keski-ikä oli 47,6 vuotta ja 65 % osallistujista oli naisia. Suurimmalla osalla (88 %) henkilöistä oli lääketieteellisen diagnoosin pääluokkaan 'M' (Tuki- ja liikuntaelimistön sairaus) kuuluva diagnoosi. useimmiten 'M54' selkäkipu (39 %) sekä 'M79' muu pehmytkudossairaus (10 %).

Poikkileikkausaineistosta analysoitiin WHODAS 2.0 kokonaispistemäärä ja tarkasteltiin kokonaispistemäärän jakaumaa visuaalisesti ja prosentuaalisesti, selvittäen mittarin lattia- ja kattovaikutus. Lisäksi osiovaste-teorian ja logistisen regressioanalyysiin avulla selvitettiin mittarin eri osioiden muuttumattomuus sukupuolen mukaan. Näiden ohella selvitettiin mittarin eri osioiden vaikeus sekä osioiden erottelukyky. Kolmanneksi selvitettiin mittarin antaman kokonaispistemäärän jakaumaan perustuen mittarin antaman kokonaispistemäärän pienin kliinisesti tärkeä ero ja pienin havaittava muutos.

Tämän väitöskirjatutkimuksen tuloksena selvisi, että kroonisilta tuki- ja liikuntaelinkipupotilailta kerätyllä aineistolla 12 kysymyksen WHODAS 2.0 -mittarin kokonaispistemäärä sekä sen kaikki osiot osoittivat merkitsevää lattiavaikutusta. Tämä tarkoittaa sitä, että mikäli kuntoutujalla on lievä toimintakyvyn rajoite, mittarin asteikon matalampi pää ei kykene erottelemaan ihmisiä joiden toimintakyky poikkeaa toisistaan. Tämän lisäksi voidaan todeta, että mittarin kahdestatoista osiosta seitsemän toimii eri tavoin miehillä ja naisilla, vaikka miesten ja naisten toimintakyvyn kokonaispistemäärä tutkitussa aineistossa oli sama. Tämä tarkoittaa sitä, että osiosta riippuen, miesten ja naisten oli koettava erilaista toimintakyvyn rajoitetta antaakseen saman vastauksen kyseessä olevaan osioon. Edelleen selvisi, että mittarin pienin havaittu muutos oli 8,6 pistettä ja mittarin kliinisesti tärkeä ero vaihteli laskukaavasta riippuen kolmen ja viiden pisteen välillä. Tämä tarkoittaa sitä, että selvitettäessä kuntoutusintervention vaikutusta toimintakykyyn 12 kysymyksen WHODAS 2.0 -mittarilla, vasta yhdeksän pisteen muutos kokonaispistemäärässä voidaan katsoa todelliseksi toimintakyvyn muutokseksi, ei mittavirheeksi.

Huomioiden edellä kuvatut kokonaispistemäärän käyttämiseen ja tulkintaan liittyvät rajoitukset, 12-osaista WHODAS 2.0 -mittaria voidaan käyttää työkykyarviointiin sekä kuntoutustarpeen ja kuntoutuksen vaikutusten arviointiin kroonisilla tuki– ja liikuntaelinkipupotilailla erityisesti, jos halutaan erotella lieviä ja vaikeita toimintakykyrajoitteita kokevat kuntoutujat toisistaan.

REFERENCES

Aalto,A., (2011). Psyykkisen toimintakyvyn mittaaminen väestötutkimuksissa. Retrieved 12.10.2020, from http://urn.fi/URN:NBN:fife201703315910

Andrews, G., Kemp, A., Sunderland, M., Von Korff, M. & Ustun, B. (2009). Normative data for the 12 item WHO disability assessment schedule 2.0. PloS one (4) 12, e8343 DOI: 10.1371/journal.pone.0008343.

Axelsson, E., Lindsater, E., Ljotsson, B., Andersson, E. & Hedman-Lagerlof, E. (2017). The 12-item self-report World Health Organization Disability Assessment Schedule (WHODAS) 2.0 Administered via the internet to individuals with anxiety and stress disorders: A psychometric investigation based on data from two clinical trials. Journal of Medical Internet Research Mental health (8) 4, e58 DOI: 10.2196/mental.7497.

Baker, F. (2001). The basis of item response theory. USA: ERIC Clearinghouse on Assessment and Evaluation.

Baron, M., Schieir, O., Hudson, M., Steele, R., Kolahi, S., et al. (2008). The clinimetric properties of the World Health Organization disability assessment schedule II in early inflammatory arthritis. Arthritis Care & Research (59) 3, 382-390 DOI: 10.1002/art.23314.

Beaton, D., Bombardier, C., Katz, J. & Wright, J. (2001). A taxonomy for responsiveness. Journal of Clinical Epidemiology (54)12, 1204-1217.

Briggs, A., Woolf, A., Dreinhöfer, Homb, N. , Hoy, D., et al. (2018). Reducing the global burden of musculoskeletal conditions. Bulletin of the World Health Organization (96) 366–368 DOI: 10.2471/BLT.17.204891.

Buist-Bouwman, M., Ormel, J., De Graaf, R., Vilagut, G., Alonso, J., et al. (2008). Psychometric properties of the World Health Organization Disability Assessment Schedule used in the European study of the epidemiology of mental disorders. International journal of methods in psychiatric research (17) 4, 185-197.

Carey, M., Asbury, J. & Tolich, M. (2012). Focus Group Research. London, United Kingdom: Taylor & Francis Group.

Carlozzi, N., Kratz, A., Downing, N., Goodnight, S., Miner, J., et al. (2015). Validity of the 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) in individuals with Huntington disease (HD). Quality of Life Research (24) 8, 1963-1971.

Cella, D., Gershon, R., Jin-Shei, L. & Seung, C. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Quality of Life Research (16) 1, 133-141.

Chatterji, S., Ustun, B. & Trotter II, R. (2001). Objectives and Overall Plan for the ICIDH-2 Cross-Cultural Applicability Research Study. In Ustun, B., Chatterji, S., Bickenbach, J., Trotter II, R., Room, R., Rehm, J. and Saxena, S. (Eds.) Disability and Culture. Universalism and Diversity.Seattle: Hogrefe & Huber Publishers, 21-27

Chisolm, T., Abrams, H., Mcardle, R., Wilson, R. & Doyle, P. (2005). The WHO-DAS II: Psychometric properties in the measurement of functional health status in adults with acquired hearingloss. Trends in Amplification (9) 3, 111-126.

Chwastiak, L. & Von Korff, M. (2003). Disability in depression and back pain: Evaluation of the World Health Organization Disability Assessment Schedule (WHO DAS II) in a primary care setting. Journal of Clinical Epidemiology (56) 6, 507-514.

Ćwirlej-Sozańska, A., Bejer, A., Wiśniowska-Szurlej, A., Wilmowska-Pietruszyńska, A., De Sire, A., et al. (2020). Psychometric properties of the Polish version of the 36-Item WHODAS 2.0 in patients with low back pain. International Journal of Environmental research and Public Health 6;17(19). DOI: 10.3390/ijerph17197284

De Vet, H. & Terwee, C. (2010). The minimal detectable change should not replace the minimal important difference. Journal of Clinical Epidemiology (63) 7, 804-805.

De Vet, H., Terwee, C., Mokkink, L. & Knol, D. (2011). Measurement in Medicine : A Practical Guide. Cambridge: Cambridge University Press.

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., et al. (2000). CORE: Clinical Outcomes in Routine Evaluation. Journal of Mental Health (9) 3, 247-255.

Federici, S., Bracalenti, M., Meloni, F. & Luciano, J. (2017). World Health Organization disability assessment schedule 2.0: An international systematic review. Disability and Rehabilitation (39) 23, 2347-2380.

Federici, S. , Meloni, F., Mancini, A., Lauriola, M. & Olivetti, B. (2009a). World Health Organisation disability assessment schedule II: contribution to the Italian validation. Disability and Rehabilitation (31) 7, 553-564.

Federici, S., Meloni, F., Lo Presti, A. (2009b). International Literature Review on WHODAS II (World Health Organization disability assessment schedule II). Life Span and Disability (12) 1, 27.

Gaskin, C., Lambert, S., Bowe, S. & Orellana, L. (2017). 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 (17) 40 DOI: 10.1186/s12874-017-0309-5.

Gómez-Benito, J., Sireci, S., Padilla, J., Hidalgo, M. & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. Psicothema (30) 1, 104-109.

Gomez-Olive, F. , Schröders, J., Aboderin, I., Byass, P., Chatterji, S., et al. (2017) Variations in disability and quality of life with age and sex between eight lower income and middle-income countries: data from the INDEPTH WHO-SAGE collaboration. BMJ Global Health (2) e000508 DOI: 10.1136/bmjgh-2017-000508.

Gray, D. & Hendershot, G. (2000). The ICIDH-2: Developments for a new era of outcomes research. Archives of Physical Medicine and Rehabilitation (81) 2, 10-14.

Guilera, G., Gómez-Benito, J., Pino, O., Rojo, J. E., Cuesta, M. J., et al. (2012). Utility of the World Health Organization Disability Assessment Schedule II in schizophrenia. Schizophrenia Research (138) 2-3, 240-247.

Guyatt, G., Kirshner, B. & Jaeschke, R. (1992). Measuring health status: What are the necessary measurement properties? Journal of Clinical Epidemiology (45) 12, 1341-1345.

Hak, T., Van Der Veer, K. & Ommundsen, R. (2006). An application of the Three ‐ Step Test ‐ Interview (TSTI): A validation Study of the Dutch and Norwegian versions of the ‘Illegal Aliens Scale. International Journal of Social Research Methodology 9(3), 215-227.

Haley, S. & Fragala-Pinkham, M. (2006). Interpreting change scores of tests and measures used in physical therapy. Physical Therapy (86) 5, 735-743.

ICF Research Branch. (2017). ICF Research Branch/ICF Core Sets. Retrieved 5.4.2020, from https://www.icf-research-branch.org/icf-core-sets.

Jablensky, A., Schwarz, R. & Tomov, T. (1980). WHO collaborative study on impairments and disabilities associated with schizophrenic disorders: A preliminary communications and methods. Acta Psychiatrica Scandinavica (62) 285, 152-163.

Kean, J. & Reilly, J. (2015). Item Response Theory. In Hammond, F., Malec, J., Buschbacher, R. and Nick, T. (Eds.) Handbook for Clinical Research : Design, Statistics, and Implementation. New York, NY: Demos Medical Publishing.

Kela (2020a). Kela kuntoutuksen palvelukuvaus. TULES-avokurssi. Retrieved 6.10.2019, from https://www.kela.fi/documents/10180/24972165/Tules+avokurssi.pdf/44e14590-13d1-4b7e-b14a-078c83a6d60b.

Kela (2020b). Kelan kuntoutuksen palvelukuvaus. TULES-kurssi. Retrieved 6.10.2019 from https://www.kela.fi/documents/10180/24972165/Tules+kuntoutuskurssi.pdf/9553655b-2b50-4a5c-b768-351cf2bc9dc2.

Kela (2020c). Kelan kuntoutuksen palvelukuvaus. Yleinen osa. Retrieved 6.10.2019 from https://www.kela.fi/documents/10180/24972165/Yleinen+osa.pdf/2024d7cf-97cd-4895-b6a1-7acbea77ddd9.

Kirchberger, I., Braitmayer, K., Coenen, M., Oberhauser, C. & Meisinger, C. (2014) Feasibility and psychometric properties of the German 12-item WHO Disability Assessment Schedule (WHODAS 2.0) in a population-based sample of patients with myocardial infarction from the MONICA/KORA myocardial infarction registry. Population Health Metrics (12) 27, DOI: 10.1186/s12963-014-0027-8.

Kohn, C., Sidovar, M., Kaur, K., Zhu, Y. & Coleman, C. (2014). Estimating a minimal clinically important difference for the EuroQol 5-Dimension health status index in persons with multiple sclerosis. Health and Quality of Life Outcomes (12) 66, DOI: 10.1186/1477-7525-12-66

Kostanjsek, N. (2011) Use of The International Classification of Functioning, Disability and Health (ICF) as a conceptual framework and common language for disability statistics and health information systems. BMC Public Health 11, S3 DOI: 10.1186/1471-2458-11-S4-S3.

Kucukdeveci, A., Oztuna, D., Elhan, A. , Kutlay , S., Yildizlar, D. , et al. (2010). Factorial structure of the World Health Organization Disability Assessment Schedule (WHODAS-11) in stroke. Clinical Rehabilitation (24) 3, 276-287.

Kutlay, Ş., Küçükdeveci, A., Elhan, A., Öztuna, D., Koç, N., et al. (2011). Validation of the World Health Organization Disability Assessment Schedule II (WHODAS-II) in patients with osteoarthritis. Rheumatology International (31) 3, 339-346.

Leonardi, M., Bickenbach, J., Ustun, B., Kostanjsek, N. & Chatterji, S. (2006). The definition of disability: what is in a name? The Lancet (368) 9543, 1219-1221.

Luciano, J. , Ayuso-Mateos, J. , Fernandez, A., Aguado, J., Serrano-Blanco, A., et al. (2010a). Utility of the twelve-item World Health Organization Disability Assessment Schedule II (WHO-DAS II) for discriminating depression "caseness" and severity in Spanish primary care patients. Quality of Life Research (19) 1, 97-101.

Luciano, J. , Ayuso-Mateos, J. , Fernández, A., Serrano-Blanco, A., Roca, M., et al. (2010b). Psychometric properties of the twelve item World Health Organization Disability Assessment Schedule II (WHO-DAS II) in Spanish primary care patients with a first major depressive episode. Journal of Affective Disorders (121)1-2, 52-58.

Luciano, J., Ayuso-Mateos, J. , Aguado, J., Fernandez, A., Serrano-Blanco, A., et al. (2010c). The 12-item World Health Organization Disability Assessment Schedule II (WHO-DAS II): a nonparametric item response analysis. BMC Medical Research Methodology (20) 10, DOI: 10.1186/1471-2288-10-45

Meesters, J. , Verhoef, J., Liem, I. , Putter, H. & Vliet Vlieland, T. (2010). Validity and responsiveness of the World Health Organization Disability Assessment Schedule II to assess disability in rheumatoid arthritis patients. Rheumatology (49) 2, 326-333.

Miller, G. (1994) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review (101) 2, 343-352 DOI: 10.1037/0033-295X.101.2.343.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. The British Medical Journal (339) 1, DOI: 10.1136/bmj.b2535.

Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., et al. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology (63) 7, 737-745.

Mokkink, Lidwine B., Terwee, Caroline B., Patrick, Donald L., Alonso, Jordi, Stratford, Paul W., et al. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research (19) 4, 539-549.

Moses, T., Miao, J. & Dorans, N. (2010). A Comparison of strategies for estimating conditional DIF. Journal of Educational and Behavioral Statistics (35) 6, 726-743.

Naughton, N. & Algar, L. (2019). Linking commonly used hand therapy outcome measures to individual areas of the International Classification of Functioning: A systematic review. Journal of Hand Therapy (32) 2, 243-261.

Novak, S., Colpe, L., Barker, P. & Gfroerer, J. (2010). Development of a brief mental health impairment scale using a nationally representative sample in the USA. International Journal of methods in psychiatric research (19) 1, 49-60.

Oliveira, C., Lee, A., Granger, C., Miller, K., Irving, L., et al. (2013). Postural control and fear of falling assessment in people with chronic obstructive pulmonary disease: a systematic review of instruments, international classification of functioning, disability and health linkage, and measurement properties. Archieves of Physical Medicine and Rehabilitation (94) 9, 1784-1799.

Paltamaa J. (2014). WHODAS 2.0 : terveyden ja toimintarajoitteiden arvioinnin käsikirjan osat 2 & 3. Jyväskylän ammattikorkeakoulun julkaisuja, Jyväskylän ammattikorkekoulu: Suomen yliopistopaino- Juvenes Print

Paltamaa, J. & Anttila, H. (2015). WHODAS 2.0 - terveyden ja toimintarajoitteiden arviointi. Toimia mittarit. Retrieved 27.6.2020 from https://www.terveysportti.fi/dtk/tmi/koti.

Paltamaa, J. & Kantanen, M. (2013). Suositus osallistumisen yleisluonteisista arviointimenetelmistä aikuisilla. Retrieved 27.6.2020, from https://www.terveysportti.fi/dtk/tmi/koti.

Prieto, L., Epping-Jordan, J., Doyle, P., Chatterji, S. & Ustun, B (2000). The factor structure of the World Health Organization Disability Assessment Schedule (WHODAS II). Quality of Life Research (9) 3, 320

Ptyushkin, P., Selb, M. & Cieza, A. (2012). ICF Core Sets. In Bickenbach, J., Cieza, A., Rauch, A. and Stucki, G. (Eds.) ICF Core Sets. Manual for clinical practice. Göttingen: Hogrefe Publishing.

Pösl, M., Cieza, A. & Stucki, G. (2007). Psychometric properties of the WHODASII in rehabilitation patients. Quality of Life Research (16) 9, 1521-1531.

Redelmeier, D., Guyatt, G. & Goldstein, R. (1996). Assessing the minimal important difference in symptoms: A comparison of two techniques. Journal of Clinical Epidemiology (49) 11, 1215-1219.

Redelmeier, D. & Lorig, K. (1993). Assessing the Clinical Importance of Symptomatic Improvements: An Illustration in Rheumatology. Archives of Internal Medicine (153) 11, 1337-1342.

Room, R., Janca, A., Bennet, L., Schmidt, L. & Sartorius, N. (1996). WHO cross-cultural applicability research on diagnosis and assessment of substance use disorders: an overview of methods and selected results. Addiction (91) 2, 199-220.

Saltychev, M., Bärlund, E., Mattie, R., Mccormick, Z., Paltamaa, J., et al. (2017a). 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 (31) 2, 262-272.

Saltychev, M., Mattie, R., Mccormick, Z. & Laimi, K. (2017b). Confirmatory factor analysis of 12-Item World Health Organization Disability Assessment Schedule in patients with musculoskeletal pain conditions. Clinical Rehabilitation (31) 5, 702-709.

Salvador-Carulla, L. & Garcia-Gutierrez, C. (2011) The WHO construct of health-related functioning (HrF) and its implications for health policy. BMC Public Health 11, S9 DOI: 10.1186/1471-2458-11-S4-S9.

Satariano, W., Guralnik, J., Jackson, R., Marottoli, R., Phelan, E., et al. (2012). Mobility and aging: new directions for public health action. American Journal Public Health (102) 8, 1508-1515.

Schneider, M., Baron, E., Davies, T., Bass, J. & Lund, C. (2015). Making assessment locally relevant: measuring functioning for maternal depression in Khayelitsha, Cape Town. Social Psychiatry and Psychiatric Epidemiology (50) 5, 797-806.

Silva, A., Cerqueira, M., Raquel Santos, A., Ferreira, C., Alvarelhao, J., et al. (2019). 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 (41) 3, 366-373.

Sloan, J., Aaronson, N., Cappelleri, J. & Fairclough, D. (2002). Assessing the clinical significance of single items relative to summated scores. Mayo Clinic Proceedings (77) 5, 479-487.

Sosiaali - Ja Terveysministeriö (2017). Kuntoutuksen uudistamiskomitean ehdotukset kuntoutusjärjestelmän uudistamiseksi. Sosiaali- ja terveysministeriön raportteja ja muistioita 2017:41. Helsinki, Sosiaali- ja terveysministeriö.

Sousa, R., Dewey, M., Acosta, D., Jotheeswaran, A., Castro-Costa, E., et al. (2010). Measuring disability across cultures--the psychometric properties of the WHODAS II in older people from seven low- and middle-income countries. The 10/66 dementia research group population-based survey. International Journal of Methods in Psychiatric research (19) 1, 1-17.

Stedman, T., Yellowlees, P., Mellsop, G., Clarke, R., Drake, S. (1997). Measuring Consumer Outcomes in Mental Health. Canberra, ACT: Department of Health and Family Services.

Stucki, G., Kostanjsek, N., Ustun, B. & Cieza, A. (2008). ICF-based classification and measurement of functioning. European Journal of Physical Rehabilitation Medicine (44) 3, 315-328.

Swaminathan, H., & Rogers, H-J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement (27) 4, 361-370.

Tarvonen-Schröder, S., Tenovuo, O., Kaljonen, A. & Laimi, K. (2018) Usability of World Health Organization Disability Assessment Schedule in chronic traumatic brain injury. Journal of Rehabilitation Medicine 50, 514-518 DOI: 10.2340/16501977-2345.

Thara, R., Rajkumar, S. & Valecha, V. (1988). The schedule for assessment of psychiatric disability – a modification of the DAS-II. Indian Journal of Psychiatry 30, 47-53.

THL (2019a). Lainsäädäntö edellyttää toimintakyvyn arviointia. Retrieved 30.9.2019, from https://thl.fi/fi/web/toimintakyky/toimintakyvyn arviointi/lainsaadanto-edellyttaa-laadukasta-arviointia.

THL (2019b). Mitä toimintakyky on? Retrieved 27.9.2019, from https://thl.fi/fi/web/toimintakyky/mita-toimintakyky-on.

THL (2019c). TOIMIA Functioning Measures Database. Retrieved 20.2.2020, from https://thl.fi/en/web/functioning/toimia-functioning-measures-database.

Thorpe, G. & Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. Psychology Faculty Scholarship.20. Retrieved 12.10.2020 from https://digitalcommons.library.umaine.edu/psy_facpub/20.

Tiikkainen, P. & Pynnönen, K. (2018) Sosiaalisen toimintakyvyn arviointi ja mittaaminen väestötutkimuksissa. Retrieved 13.10.2020 from http://urn.fi/URN:NBN:fife201703315912.

Trotter, R., Ustun, B., Chatterji, S., Rehm, J., Room, R., et al. (2001). Cross-cultural applicability research on disablement: Models and methods for the revision of an international classification. Human Organization (60) 1, 13-27.

Tuulio-Henriksson, A. (2011) Kognitiivisen toimintakyvyn arviointi väestötutkimuksissa. Retrieved 13.10.2020 from http://urn.fi/URN:NBN:fi-fe201703315911

United Nations (2006). Convention on the rights of persons with disabilities, UN General Assembly.

Ustun, B., Rehm, J., Chatterji, S., Saxena, S., Trotter, R., et al. (1999). Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. The Lancet (354) 9173, 111-115.

Ustun, B., Chatterji, S., Bickenbach, J., Trotter II, R., Room, R., et al. (2001). Cross-Cultural Results. Summary and conclusions. In Ustun, B., Chatterji, S., Bickenbach, J., Trotter II, R., Room, R., Rehm, J. and Saxena, S. (Eds.) Disability and culture. Universalism and Diversity. Seattle: Hofgre & Huber Publishers: 320

Ustun, B., Chatterji, S., Bickenbach, J., Trotter II, R. & Room, R. (2003b). Extended review. Disability & Society (18) 6, 827-833.

Ustun, B., Chatterji S., Mechbal, A. & Murray, C. (2003c). The World Health Surveys In Murray, C. & Evans, D. (Eds.) Health systems performance assessment : debates, methods and empiricism. Geneva: World Health Organization: 797-806.

Ustun, B., Chatterji, S., Villanueva, M., Bendib, L., Çelik, C., et al. (2003a). WHO Multi-country Survey Study on Health and Responsiveness 2000–2001. In Murray, C. and Evans, D. (Eds.) Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization.

Ustun, B., Chatterji, S., Kostanjsek, N., Rehm, J., Kennedy, C., et al. (2010a). Developing the World Health Organization Disability Assessment Schedule 2.0. Bulletin of the World Health Organization (88) 11, 815-823.

Ustun, B. Kostanjesek, N. Chatterji, S. & Rehm, J. Eds. (2010b). Measuring health and disability : manual for WHO Disability Assessment Schedule (WHODAS 2.0). Geneva: World Health Organization.

Van Der Zee, C., Post, M., Brinkhof, M. & Wagenaar, R. (2014). Comparison of the Utrecht scale for evaluation of rehabilitation-participation with the ICF Measure of participation and activities Screener and the WHO Disability Assessment Schedule II in persons with spinal cord injury. Archieves of Physical Medicine and Rehabilitation (95) 1, 87-93.

Wasiak, J., Mcmahon, M., Danilla, S., Spinks, A., Cleland, H., et al. (2011). Measuring common outcome measures and their concepts using the International Classification of Functioning, Disability and Health (ICF) in adults with burn injury: a systematic review. Burns (37) 6, 913-924.

WHO (1980). International classification of impairments, disabilities, and handicaps : a manual of classification relating to the consequences of disease. Geneva: World Health Organization.

WHO (1988). WHO psychiatric disability assessment schedule (WHO/DAS) : with a guide to its use. Geneva: World Health Organization.

WHO (2001). International classification of functioning, disability and health : ICF. Geneva: World Health Organization.

WHO (2003). ICF Checlist. Retrieved 4.4.2020 from https://www.who.int/classifications/icf/icfchecklist.pdf?ua=1.

WHO (2019). International statistical classification of diseases and related health problems (11th ed.). Retriewed 8.1.2021 from https://icd.who.int/.

Wolf, A., Tate, R., Lannin, N., Middleton, J., Lane-Brown, A., et al. (2012). 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. Journal of Rehabilitation Medicine (44) 9, 747-755.

Von Korff, M., Crane, P., Alonso, J., Vilagut, G., Angermeyer, M., et al. (2008). Modified WHODAS-II provides valid measure of global disability but filter items increased skewness. Journal of Clinical Epidemiology (61) 11, 1132-1143.

Yoon, J. , Jung, M., Shin, I., Yang, S., Zheng, T, et al. (2004). Development of Korean version of World Health Organization Disability Assessment Schedule II (WHODAS II-K) in Community Dwelling Elders. Journal of Korean Neuropsychiatric Association (43) 1, 86-92.

Yu, C. (2017). A Simple Guide to the item response theory (IRT) and Rasch Modeling.   Retrieved 26.4.2020, from http://www.creative-wisdom.com/computer/sas/sas.html.

Ziebland, S., Fitzpatrick, R. & Jenkinson, C. (1993). Tacit models of disability underlying health status instruments. Social Science & Medicine (37) 1, 69-75.

Zumbo, B., & Gelin, M. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. Journal of Educational Research & Policy Studies (5) 1, 1-23.

Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly (4) 2, 223-233.

ORIGINAL PAPERS

I

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

by

Mikhail Saltychev, Niina Katajapuu, Esa Bärlund & Katri Laimi 2019

Disability and Rehabilitation vol 23, 1-6

DOI: 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

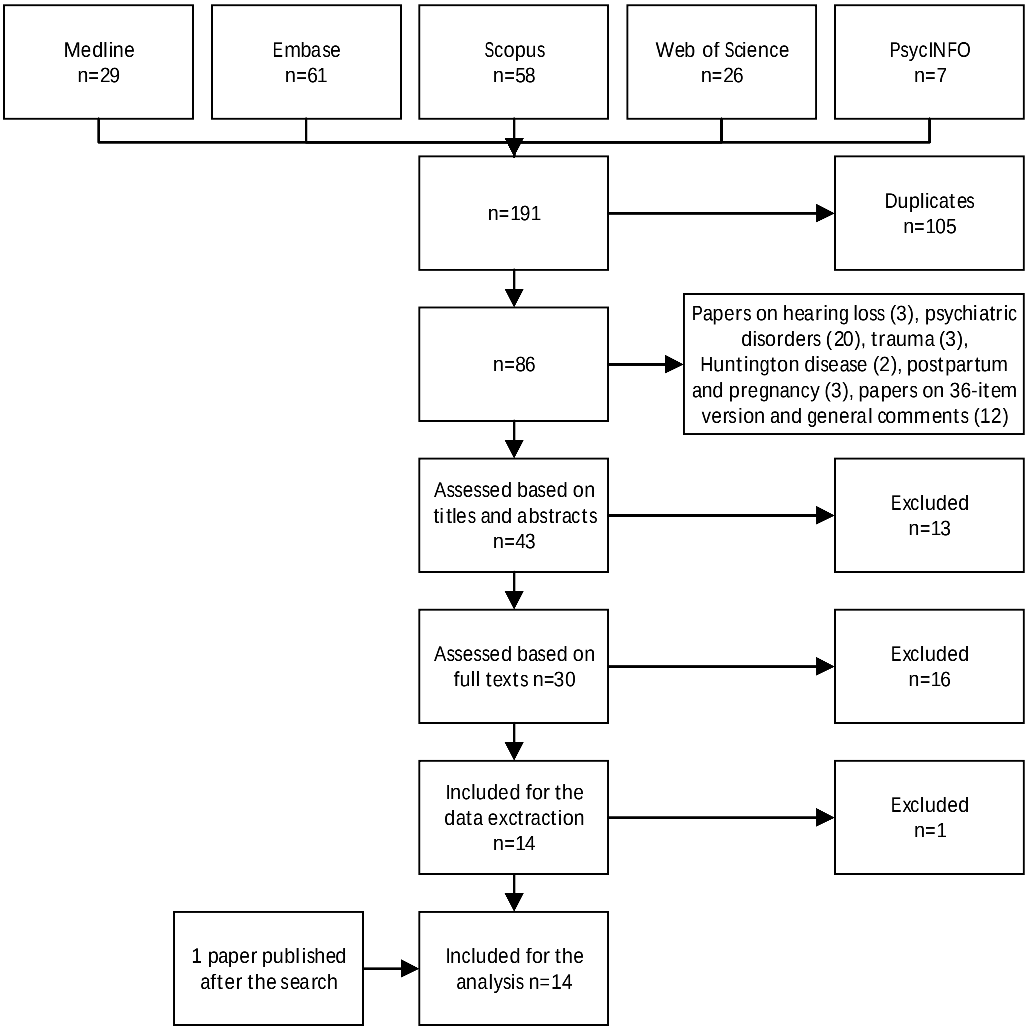
Table 2. Basic characteristics of the included studies

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


Medline n=29
Embase n=61
Scopus n=58
Web of Science n=26
PsycINFO n=7
n=191
Duplicates n=105
n=86
Papers on hearing loss (3), psychiatric disorders (20), trauma (3), Huntington disease (2), postpartum and pregnancy (3), papers on 36-item version and general comments (12)
Assessed based on titles and abstracts n=43
Excluded n=13
Assessed based on full texts n=30
Excluded n=16
Included for the data exctraction n=14
Excluded n=1
1 paper published after the search
Included for the analysis n=14

II

FLOOR AND CEILING EFFECTS OF THE WORLD HEALTH ORGANIZATION DISABILITY ASSESSMENT SCHEDULE 2.0 AMONG PATIENTS WITH CHRONIC MUSCULOSKELETAL PAIN

by

Niina Katajapuu, Katri Laimi, Ari Heinonen & Mikhail Saltychev 2019

International Journal of Rehabilitation Research vol 42, 190–192

DOI 10.1097/MRR.0000000000000339

Floor and ceiling effects of the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) amongst patients with chronic musculoskeletal pain

Short title: **WHODAS 2.0 floor and ceiling effects**

Niina Katajapuu MSc[1], Katri Laimi PhD, MD[2], Ari Heinonen PhD[1], Mikhail Saltychev PhD, MD[2]

[1] Faculty of Health and Sport Sciences, University of Jyväskylä, Jyväskylä, Finland

[2] Department of Physical and Rehabilitation Medicine, Turku University Hospital and University of Turku, Turku, Finland.

ADDRESS FOR CORRESPONDENCE

Niina Katajapuu, Email: niina.katajapuu@turkuamk.fi

Turku University of Applied Sciences

Joukahaisenkatu 3, 20520 Turku, Finland

CONFLICT OF INTEREST

None declared

FUNDING

None declared

ACKNOWLEDGEMENTS

None declared

ABSTRACT

Objective of this study was to investigate the floor and ceiling effects of 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS). This was a cross-sectional survey study at a university's Physical and Rehabilitation Medicine (PRM) outpatient clinic amongst 1988 patients with chronic musculoskeletal pain. Floor and ceiling effects were calculated as relative frequencies of the lowest or the highest possible scores for each item. Probit plotting method was used to detect the non-normality of distribution of total score graphically. A significant floor effect of 15% to 79% was observed in all twelve WHODAS 2.0 items. A substantial floor effect for total score was detected as well graphically. No ceiling effects were observed. In this study, significant floor effect was found for all WHODAS 2.0 items amongst patients with chronic musculoskeletal pain associated with mild or no disability.

KEYWORDS

Musculoskeletal pain, WHODAS 2.0, floor effect, ceiling effect, validity

INTRODUCTION

In an ideal situation, a scale is able to measure the entire spectrum of a phenomenon. However, scales commonly perform better around their mid area displaying a poorer discrimination ability at their tails producing so called floor and ceiling effects. Statistically speaking, 'floor effect' is a level below and 'ceiling effect' a level above which variance within an independent variable is no longer measurable (Velozo et al., 2012, De Vet, 2011). For example, when using a pain numeric rating scale, estimates might cluster around zero demonstrating a significant floor effect in a sample predominated by patients with mild pain severity or without pain. In other words, in that hypothetical case, a numeric rating scale may fail to distinguish people with very mild pain from those with no pain at all – both will mark zero.

Floor and ceiling effects are common findings when measuring functioning restrictions amongst people with musculoskeletal disorders (McHorney et al., 1994, Pellicciari et al., 2016). Modest to substantial ceiling effects have been found for the 36-item Short Form Health Survey (SF -36) in chronic medical and psychiatric conditions (McHorney et al., 1994). Small average scores of Neck Disability Index in patients with acute neck pain have probably been related to a floor effect in primary care population (Vos et al., 2006). A 36% floor effect of Neck Disability Index has also been observed in patients with neck pain in a university spinal clinic (Hung et al., 2015). Floor effects of Daily Activity Questionnaire have been detected in patients with different musculoskeletal conditions, the worst floor effect being observed in ankylosing spondylitis and Sjögren syndrome (Hammond et al., 2018).

WHODAS 2.0. is a generic tool to assess health and disability across all diseases and cultures in both clinical and general population settings (World Health Organization, 2018, Chiu et al., 2014, Carlozzi et al., 2015, Younus et al., 2017). While the WHODAS 2.0 psychometrics have extensively been studied, only a few inconsistent reports on its floor and ceiling effects have been published so far. Federici et al. have observed strong 75% floor effect in 'self-care' and 60% in 'getting around' domains amongst healthy volunteers and, respectively, 50% and 40% amongst disabled patients (Federici et al., 2009). Significant floor effects in "understanding and communicating" and "getting along with people" domains and a milder floor effect in 'self-care' domain amongst patients with rheumatoid arthritis have been reported (Meesters et al., 2010). In turn, when studying patients with spinal cord injury, a floor effect has not been observed but, instead, a large 54% ceiling effect in “understanding and communicating” domain and milder ceiling effects in “self-care” and “getting along with others” domains (Wolf et al., 2012). Similar results amongst patients with spinal cord injury have been seen: significant ceiling effects in items “understanding and communication”, “self-care”, and “getting along with others” (van der Zee et al., 2014). The recent review has reported a floor effect within ”self-care” domain explaining the finding by cultural differences (Federici et al., 2017). Both floor and ceiling effects across most of the domains of a modified 36-item WHODAS 2.0 have been observed (Yen et al., 2014). Previous research has suggested further evaluation of floor and ceiling effects of WHODAS 2.0 within different patient groups. Knowledge on how well WHODAS 2.0 performs across the entire spectrum of restricted functioning amongst patients with chronic musculoskeletal pain may improve its usability in different situations as e.g., screening, clinical evaluation, or attaining rehabilitation goals. The aim of this study was to assess the ceiling and floor effects of WHODAS 2.0 amongst patients with chronic musculoskeletal pain.

METHODS

This was a cross-sectional study of consecutive patients with chronic musculoskeletal pain who were seen in an outpatient Physical and Rehabilitation Medicine (PRM) clinic of university hospital between April 2014 and February 2017. The survey was sent to the patients and filled up before a physician appointment. The survey included the WHODAS 2.0 questionnaire and questions on demographics, pain intensity, perceived general health, and working ability among others. A university hospital ethics committee approved the study.

The self-administered WHODAS 2.0 contains 12 items covering the most common limitations of functioning appearing in general population. The questionnaire covers limitations during the last 30 days. A Likert-type scale is used to define the severity of limitation with 0 denoting "no limitation" and 4 denoting "extreme limitation or inability to function". For the calculations employed in this study, the total score was the sum of all 12 responses divided by 48 and multiplied by 100 and presented as a percentage where 100% represents the worst possible restriction.

Age was defined in full years at the time of visiting the clinic. Pain intensity was assessed using an 11-point numeric rating scale (NRS) 0 denoting "no pain" and 10 denoting "worst possible pain". Educational level was dichotomized "high school" vs. "no high school". Body mass index (BMI) was calculated as a body mass divided by a squared body height (kg/m^2).

Statistical analysis

The basic characteristics were presented as means, standard deviations (SDs), and percentage when appropriate. In case of a rough 5-point Likert-type scale used in WHODAS 2.0 individual items, the ceiling and floor effects of WHODAS 2.0 were calculated numerically as a relative frequency of lowest or highest possible score achieved by the respondents (McHorney et al., 1994, Coster et al., 2014, Carlozzi et al., 2015). The cut-off for a significant floor or ceiling effect was set at >=15%. Instead, the distribution of a continuous WHODAS 2.0 total score was analyzed graphically. To detect the nonnormality of WHODAS 2.0 total score's distribution, the probit plotting method was used as described by Miller.(Miller R.G, 1997) This method demonstrates how the sample is differing from normality and presents irregularities in the tails rather than only in the middle of the distribution. All the analyses were conducted using Stata/IC Statistical Software: Release 15. College Station (StataCorp LP, TX, USA).

RESULTS

Of 3150 patients visiting the clinic, 1988 (63%) returned a questionnaire. The patients were 47.6 (15.0) year-old and 1,297 (65%) were women (Table 1). The average intensity of pain was 6.3 (2.0) points. Most of the patients (n=1746, 88%) had a main diagnosis 'M' - 'Diseases of the musculoskeletal system and connective tissue' - according to the International Classification of Diseases version 10. The most frequent diagnoses were 'M54 Dorsalgia' (n=781, 39%) and 'M79 Other soft tissue disorders' (n=202, 10%). A significant floor effect was observed in all twelve WHODAS 2.0 items varying from 15% to 79% (Table 2). Figure 1 displays the substantial floor effect for a total score as well. No ceiling effect was detected for any of WHODAS 2.0 items.

DISCUSSION

This cross-sectional study amongst 2000 patients with chronic musculoskeletal pain showed a significant floor effect for all twelve items of WHODAS 2.0 and for its total score.

The sample represented a population of patients referred to a university PRM outpatient clinic supposing to receive high-end examination and treatment and, thus, they probably differ from patients treated in primary healthcare. The sample was predominated by women. As the patients experienced mostly mild disability, the generalization of the results over populations with more severe disability levels (e.g., in in-patient settings) may be problematic. However, the sample was large enough to achieve credible results for the population of interest – mildly disabled patients with chronic musculoskeletal pain.

The results were in line with some previous studies that detected floor effects for several WHODAS 2.0 items or for its overall score (Federici et al., 2009, Meesters et al., 2010, Schneider M et al., 2015). On contrary, the results differed from previously observed ceiling effect in 'self-care' domain without any significant floor effects amongst patients with spinal cord injury (Wolf et al., 2012, van der Zee et al., 2014). In other words, the results of the present and previous studies pointed the possibility that WHODAS 2.0 may have a substantial floor effect when disability is mild and ceiling effect in situations where disability is severe. This conclusion raises a question if WHODAS 2.0 is sensitive amongst people with midrange disability levels only?

Further research is needed especially amongst populations with midrange disability severity and amongst mixed samples containing patients of all grades of disability.

Amongst mildly disabled patients with chronic musculoskeletal pain, significant floor effect was found for all WHODAS 2.0 items.

REFERENCES

CARLOZZI, N. E., KRATZ, A. L., DOWNING, N. R., GOODNIGHT, S., MINER, J. A., MIGLIORE, N. & PAULSEN, J. S. 2015. Validity of the 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) in individuals with Huntington disease (HD). *Qual Life Res,* 24**,** 1963-71.

CHIU, T. Y., YEN, C. F., CHOU, C. H., LIN, J. D., HWANG, A. W., LIAO, H. F. & CHI, W. C. 2014. Development of traditional Chinese version of World Health Organization disability assessment schedule 2.0 36--item (WHODAS 2.0) in Taiwan: validity and reliability analyses. *Res Dev Disabil,* 35**,** 2812-20.

COSTER, M. C., BREMANDER, A., ROSENGREN, B. E., MAGNUSSON, H., CARLSSON, A. & KARLSSON, M. K. 2014. Validity, reliability, and responsiveness of the Self-reported Foot and Ankle Score (SEFAS) in forefoot, hindfoot, and ankle disorders. *Acta Orthop,* 85**,** 187-94.

DE VET, H., TERWEE, C., MOKKINK, L., & KNOL, D. 2011. *Measurement in Medicine: A Practical Guide (Practical Guides to Biostatistics and Epidemiology.*

FEDERICI, S., BRACALENTI, M., MELONI, F. & LUCIANO, J. V. 2017. World Health Organization disability assessment schedule 2.0: An international systematic review. *Disability and Rehabilitation,* 39**,** 2347-2380.

FEDERICI, S., MELONI, F., MANCINI, A., LAURIOLA, M. & OLIVETTI BELARDINELLI, M. 2009. World Health Organisation Disability Assessment Schedule II: contribution to the Italian validation. *Disabil Rehabil,* 31**,** 553-64.

HAMMOND, A., PRIOR, Y., HORTON, M. C., TENNANT, A. & TYSON, S. 2018. The psychometric properties of the Evaluation of Daily Activity Questionnaire in seven musculoskeletal conditions. *Disabil Rehabil,* 40**,** 2070-2080.

HUNG, M., CHENG, C., HON, S. D., FRANKLIN, J. D. & LAWRENCE, B. D. 2015. Challenging the norm: further psychometric investigation of the Neck Disability Index. *Spine J*.

MCHORNEY, C. A., WARE, J. E., JR., LU, J. F. & SHERBOURNE, C. D. 1994. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care,* 32**,** 40-66.

MEESTERS, J. J., VERHOEF, J., LIEM, I. S., PUTTER, H. & VLIET VLIELAND, T. P. 2010. Validity and responsiveness of the World Health Organization Disability Assessment Schedule II to assess disability in rheumatoid arthritis patients. *Rheumatology (Oxford),* 49**,** 326-33.

MILLER R.G 1997. *Beyond anova: Basics of applied statistics,* London, Chapman & Hall.

PELLICCIARI, L., BONETTI, F., DI FOGGIA, D., MONESI, M. & VERCELLI, S. 2016. Patient-reported outcome measures for non-specific neck pain validated in the Italian-language: a systematic review. *Archives of Physiotherapy,* 6**,** 9.

SCHNEIDER M, BARON E, DAVIES T, J, B. & LUND C 2015. Making assessment locally relevant: measuring functioning for maternal depression in Khayelitsha, Cape Town. *Soc Psychiatry Psychiatr Epidemiol,* 50**,** 797-806.

VAN DER ZEE, C. H., POST, M. W., BRINKHOF, M. W. & WAGENAAR, R. C. 2014. Comparison of the Utrecht Scale for Evaluation of Rehabilitation-Participation with the ICF Measure of Participation and Activities Screener and the WHO Disability Assessment Schedule II in persons with spinal cord injury. *Arch Phys Med Rehabil,* 95**,** 87-93.

VELOZO, C. A., SEEL, R. T., MAGASI, S., HEINEMANN, A. W. & ROMERO, S. 2012. Improving Measurement Methods in Rehabilitation: Core Concepts and Recommendations for Scale Development. *Archives of Physical Medicine and Rehabilitation,* 93**,** S154-S163.

WOLF, A. C., TATE, R. L., LANNIN, N. A., MIDDLETON, J., LANE-BROWN, A. & CAMERON, I. D. 2012. 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,* 44**,** 747-55.

WORLD HEALTH ORGANIZATION. 2018. *WHO Disability Assessment Schedule 2.0 (WHODAS 2.0)* [Online]. Available: http://www.who.int/classifications/icf/whodasii/en/ [Accessed 21.10. 2018].

VOS, C. J., VERHAGEN, A. P. & KOES, B. W. 2006. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. *Eur Spine J.,* 15.

YEN, C. F., HWANG, A. W., LIOU, T. H., CHIU, T. Y., HSU, H. Y., CHI, W. C., WU, T. F., CHANG, B. S., LU, S. J., LIAO, H. F., TENG, S. W. & CHIU, W. T. 2014. Validity and reliability of the Functioning Disability Evaluation Scale-Adult Version based on the WHODAS 2.0--36 items. *J Formos Med Assoc,* 113**,** 839-49.

YOUNUS, M. I., WANG, D. M., YU, F. F., FANG, H. & GUO, X. 2017. Reliability and validity of the 12-item WHODAS 2.0 in patients with Kashin-Beck disease. *Rheumatol Int,* 37**,** 1567-1573.

TABLES AND FIGURES

Figure 1. Floor and ceiling effects of WHODAS 2.0 total score

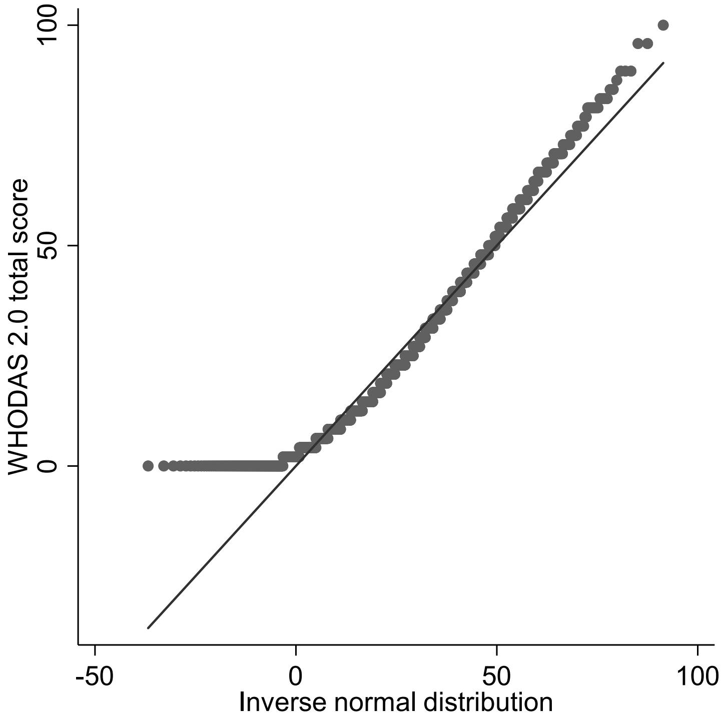


Table 1. Demographic characteristics of participants

Variable	Estimate
Age, years	47.6 (15.0)
WHODAS 2.0 total score, points	27.3 (19.5)
Body mass index, kg/m^2	27.4 (5.7)
Pain, points 0–10	6.3 (2.0)
Educational level, n	
High school	609 (33%)
No high school	1258 (67%)
Gender, n	
Women	1297 (65%)
Men	691 (35%)

Table 2. Floor and ceiling effects of WHODAS 2.0 individual items

Item	Lowest score ‘0’	Highest score ‘4’
Standing for long periods	29%	0%
Taking care of household responsibilities	21%	0%
Learning new task	74%	1%
Joining in community activities	46%	6%
Emotional affection by health problems	15%	3%
Concentrating doing something for 10 min	56%	2%
Walking long distance as 1 km	37%	14%
Washing whole body	51%	2%
Getting dressed	42%	1%
Dealing with people you do not know	79%	2%
Maintaining friendship	62%	2%
Day-to day work/school	18%	0%

III

GENDER-RELATED DIFFERENCES IN PSYCHOMETRIC PROPERTIES OF WHO DISABILITY ASSESSMENT SCHEDULE 2.0

by

Niina Katajapuu, Katri Laimi, Ari Heinonen & Mikhail Saltychev, 2019

International Journal of Rehabilitation Research vol 42, 316-321

DOI 10.1097/MRR.0000000000000365

Gender-related differences in psychometric properties of World Health Organization Disability Assessment Schedule (WHODAS 2.0)

Short title: **Differential item functioning of WHODAS 2.0**

Niina Katajapuu MSc[1,3], Katri Laimi PhD, MD[2], Ari Heinonen PhD[1], Mikhail Saltychev PhD, MD[2]

[1] Faculty of Health and Sport Sciences, University of Jyväskylä, Jyväskylä, Finland

[2] Department of Physical and Rehabilitation Medicine, Turku University Hospital and University of Turku, Turku, Finland

[3]Faculty of Health and Wellbeing, Turku University of Applied Sciences, Turku, Finland

ADDRESS FOR CORRESPONDENCE

Niina Katajapuu, Turku University of Applied Sciences, Joukahaisenkatu 3, 20540 Turku, Finland

Email: niina.katajapuu@turkuamk.fi

CONFLICT OF INTEREST

None to declare

FUNDING

None to declare

ACKNOWLEDGEMENTS

None to declare

ABSTRACT

To investigate the gender-related DIF in 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS) amongst patients with chronic musculoskeletal pain**.** Cross-sectional survey study with 1988 consecutive chronic musculoskeletal pain patients at university's Physical and Rehabilitation Medicine outpatient clinic. To assess a DIF, WHODAS 2.0 items were dichotomized as 'none' rated by respondents as '0' versus 'any limitation' rated as '1,2,3 or 4'. The item response theory analysis (IRT) was used to define discrimination and difficulty parameters of a questionnaire. The probit logistic regression was used to test uniformity of DIF between gender groups. The results of DIF analysis were presented and evaluated graphically as item characteristic curves based on 2-parameter IRT analysis of dichotomized responses. High to perfect discrimination ability was observed for all the items except one. Difficulty levels of eight items were shifted towards the elevated disability level, four items demonstrated a perfect difficulty property. Significant DIF between genders was observed in seven out of 12 items. All the detected DIFs were uniform. For item 'household', 'emotional affection' and 'work', men had to experience slightly worse disability than women to achieve the same score. A reverse effect was observed for items 'concentration', 'washing', 'dressing' and dealing with strangers. In this study, significant DIF between genders was found in seven of twelve items of 12-item WHODAS 2.0. amongst 1988 patients with chronic musculoskeletal pain. All the detected DIFs were uniform. Even if this study showed gender-related DIF in seven out of 12 items, we recommend using and studying 12-item WHODAS 2.0 in different populations.

KEYWORDS

Musculoskeletal pain, WHODAS 2.0, differential item functioning, validity

INTRODUCTION

Gender differences in the prevalence of both musculoskeletal pain and disability are well documented (Bartley. E, 2013, Bingefors and Isacson, 2004, Merril S et al., 1997). The tools to measure pain or disabilities have to be reliable in both genders. However, the reference values of many proxy-rated, patient-reported and objective outcome measures have also shown to be gender-dependent (Bohannon, 1997, Massy-Westropp et al., 2011, Dewitt R, 2013).

The 12-item version of World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) is a generic tool to assess health and disability across different diseases and cultures in both clinical settings and general population. The total score of WHODAS 2.0 has been reported to be age-dependent showing a 0.6 point-increase for every age year (Gomez-Olive et al., 2017). While other psychometric properties of WHODAS 2.0 have been studied (Carlozzi et al., 2015, Ustun et al., 2010, Saltychev et al., 2017), the occurrence of a gender-related differential item functioning (DIF) is not known. This means that we do not know, if WHODAS 2.0 is similarly sensitive in men and women with musculoskeletal pain. In two previous studies on patients with either major depression or myocardial infarction, no gender-related differences in any of WHODAS 2.0 items were observed (Kirchberger et al., 2014, Luciano et al., 2010). A study of the 36-item WHODAS 2.0 on patients with knee osteoarthritis, showed a gender-related DIF only in a 'getting housework done as quickly as needed' -item and gender-related DIF was observed in a modificated 36 -item WHODAS functioning scale off people with mental health problems (Novak et al., 2010, Kutlay et al., 2011). The 12-item WHODAS is widely used in patients with musculoskeletal symptoms, but the gender-related DIF of this scale has not been evaluated in musculoskeletal pain yet. This information is important to justify the wide use of 12-item WHODAS in screening disability, in evaluating functioning, and in planning and reviewing rehabilitation.

The aim of this study was to investigate if there is a significant gender-related DIFin a 12-item WHODAS 2.0 amongst patients with chronic musculoskeletal pain.

METHODS

This was a cross-sectional study of consecutive patients with chronic musculoskeletal pain in an outpatient Physical and Rehabilitation Medicine (PRM) clinic of a university hospital between April 2014 and February 2017. The survey was sent to the patients and filled up before a physician appointment. The survey included a 12-item WHODAS 2.0 questionnaire and questions on demographics, pain intensity, perceived general health, and working ability among others. The university hospital ethics committee approved the study.

The self-administered WHODAS 2.0 contains 12 items covering the most common limitations of activity and participation during the last 30 days. A Likert-like scale is used to define the severity of limitation from 0 to 4 with 0 denoting "no limitation" and 4 denoting "extreme limitation or inability to function". The total score was calculated in our study as a sum of all 12 items divided by 48, multiplied by 100, and presented as a percentage where 100% represents the worst possible restriction.

Age was defined in full years at the time of visiting a clinic. Pain intensity was assessed using a 11-point numeric rating scale (NRS) with 0 denoting "no pain" and 10 denoting "worst possible pain". Educational level was dichotomized "high school" vs. "no high school". Body mass index (BMI) was calculated as a body weight divided by a squared height (kg/m^2).

Statistical analysis

The basic characteristics were presented as means, standard deviations (SDs), and percentage, when appropriate. Independent t-test and chi square test were used to investigate potential differences between men and women regarding their age, educational level, BMI, and pain intensity.

Differential item functioning (DIF) is a statistical characteristic of a scale item (here counted for each of 12 items included in WHODAS 2.0) that describes if the item is measuring an ability (here level of functioning) differently for separate subgroups (here genders) within the sample. To assess a DIF, WHODAS 2.0 items were dichotomized as 'none' (rated by respondents as '0') versus 'any limitation' (rated by respondents as ‘1’, ‘2’, ‘3’, or ‘4’). It has previously been reported that such a dichotomous version of WHODAS 2.0 is compatible with its polytomous version (World Health Organisation, 2010).

The item response theory (IRT) analysis defined discrimination and difficulty parameters of a questionnaire. A discrimination parameter describes the sensitivity of test to differentiate severity levels of symptoms. The steeper the regression curve, the more discriminative the test becomes. In this study, discrimination of 0.01 to 0.24 was considered 'none' (a totally level regression curve), 0.25 to 0.64 'low', 0.65 to 1.34 'moderate', 1.35 to 1.69 'high', and a discrimination >1.7 was considered 'perfect' (a regression curve approaching a vertical line) (Baker FB, 2001). Ideally, the steepest interval corresponds to the patients who obtained average WHODAS 2.0 total scores in the studied population. In turn, difficulty is a psychometric property of a single item or an entire test, which describes how much more or less a respondent should perceive the studied ability (comparing with the average level of studied population) in order to achieve a 0.5 probability to give a particular answer.

The probit logistic regression was used to test whether an item exhibits either uniform or nonuniform DIF between gender groups that is, whether an item favors one group over the other for all values of the functioning limitation or for only some values of that (de Boeck P, 2004, Swaminathan H and Rogers HJ, 1990). A uniform DIF occurs when the difference between groups remains the same across the entire scale. In turn, a nonuniform DIF is observed when the direction of difference between groups varies at different levels of functioning limitation (e.g., if men perform better up than women to a midpoint and worse than women after that). A two-tailed *p*-value =<0.05 indicated a significant difference between genders. When significant DIF was observed, the results of DIF analysis were also presented and evaluated graphically as item characteristic curves based on 2-parameter IRT analysis of dichotomized responses.

All the analyses were conducted using Stata/IC Statistical Software: Release 15. College Station (StataCorp LP, TX, USA).

RESULTS

Of 3,150 patients visiting the clinic, 1,988 (63%) participated the study. The patients were 47.6 (SD 15.0) years old and 1,297 (65%) were women (Table 1). The average intensity of pain was 6.3 (SD 2.0) points. Most of the patients (n=1746, 88%) had a main diagnosis 'M' - 'Diseases of the musculoskeletal system and connective tissue' according to the International Classification of Diseases 10th Edition. The most frequent single diagnoses were 'M54 Dorsalgia' (n=781, 39%) and 'M79 Other soft tissue disorders' (n=202, 10%).

The total scores of WHODAS 2.0 were 27.3 (SD 19.5) points for both men and women (*p*=0.843). Probably due to a large sample size, the differences between men and women in BMI (*p*<0.001), pain severity (*p*=0.005, 95% CI 0,38 - 0.01), and educational level (*p*<0.001) were statistically significant even if the absolute estimates differed only a little.

High to perfect discrimination ability was observed for all the items except for item #9 ”dressing” with moderate discrimination (Table 2). Difficulty levels of eight items – #3, #4, #6, #7, #8, #9, #10, and #11 (learning, joining in community, concentrating, walking, washing, dressing, dealing with strangers, maintaining friendships) – were shifted towards the elevated disability level compared to average disability level of the entire studied population. In other words, musculoskeletal patients with mild or none disability clustered around the lowest possible scores on these items. Other four items (standing, household responsibilities, being emotionally affected, work) demonstrated a perfect difficulty property (Table 3).

Significant DIF between genders was observed in seven out of 12 items: ‘household responsibilities’, ‘being emotionally affected’, ‘concentrating for 10 minutes’, ‘washing’, ‘dressing’, ‘dealing with strangers’, and ‘work’. All the detected DIFs were uniform (Table 4 and Figure 1). For items #2, #5 and #12 (household, emotional affection, work), men had to experience slightly worse disability than women to achieve the same score. A reverse effect was observed for items #6, #8, #9 and #10 (concentration, washing, dressing, dealing with strangers).

DISCUSSION

This study amongst 1,988 patients with chronic musculoskeletal pain showed significant DIF between genders in seven of twelve items of 12-item WHODAS 2.0. All the detected DIFs were uniform meaning that the direction of gender-related differences between responses persisted across the entire spectrum of disability severity. For items ‘household responsibilities’, ‘emotional affection’, and ‘ work’, men had to experience slightly worse disability than women to achieve the same score. A reverse effect was observed for items ‘concentrating’, ‘washing’, ‘dressing’ and ‘dealing with strangers’.

The generalizability of the results is weakened by the fact that the sample represented a population with chronic musculoskeletal pain treated in a highly specialized health care unit (university PRM clinic). Thus, the patients might differ from those treated in e.g. primary health care. Additionally, the sample was predominated by women. This was, however, the first study on gender-related DIF of the 12-item WHODAS 2.0 with a sample large enough to achieve statistically significant results and narrow confidence intervals.

The results were similar with two previous studies on the subject (Kutlay et al., 2011, Novak et al., 2010). Novak et al. found gender-related DIF in many daily activities in people with mental health problems. In turn, Kutlay et al. reported gender-related DIF in ‘life activities’ in patient with knee osteoarthritis. Both studies also observed a significant DIF in ‘taking care of household responsibilities’ as seen in the present study: compared to women, men had to experience worse disability to reach a similar score in this item. Neither of these studies used the shortest 12-item WHODAS version. Both reports differ from the present study by populations of interest and by statistical methods used. To assess DIF, Novak et al. used odds ratio. The participants in that study were asked to evaluate their functioning level based on their worse month in previous year. Opposing, the 12–item WHODAS 2.0 is based on responses concerning previous 30 days. The study by Kutlay et al employed a 36-item WHODAS 2.0. Silva et al. have reported on some gender differences in responses to the WHODAS 2.0. In that study, men stated more often that they are not doing housework marking the item ‘household work’ as ‘not applicable’. In the present study, this phenomenon was not observed probably due to cultural differences between studied populations. Study from Silva et al differed from the present study by statistical methods and WHODAS 2.0. version used. They did not assess the differential item functioning and employed a 36-item WHODAS 2.0

The findings in this study differed from the results of two previous reports that did not observe DIF in any of WHODAS items (Luciano et al., 2010, Kirchberger et al., 2014) . Differently to present study, they employed samples of patients with acute health conditions like myocardial infarction and major depressive episode and the respondents were older when compared to present study population.

The design used in this study does not provide any explanations to the gender differences in psychometric properties of 12-item WHODAS 2.0. As this is the first study on patients with chronic musculoskeletal pain, the results cannot be straightly reflected to previous studies either. If the found differences between genders in this study population were only due to the gender-specific way of reporting disability with one gender over-estimating and other underestimating functioning limitations, the gender effect would probably not change across items as in our study. Women had to experience slightly worse disability to get the same score than male in four items, while male had to have more disability to achieve the score of women in three items. This finding supports the results of a previous study in elderly (Merrill et al 1997), where self-reported disability was highly associated with measured difficulties in both genders.

Further research in needed to reveal possible gender-related DIFs in other settings and patient groups with different levels of functioning. Repeated measures design may reveal potential fluctuations in DIFs over time. In the light of our study results, men and women might answer differently to part of WHODAS 2.0 items.

Even if this study showed gender-related DIF in seven out of 12 items of the self-administered WHODAS 2.0 in musculoskeletal pain, these differences were uniform across the whole scale of severity, and we still recommend using and studying 12-item WHODAS 2.0 in different populations.

REFERENCES

BAKER FB 2001. *The basics of item reponse theory,* USA, ERIC Clearinghouse on Assessment and Evaluation.

BARTLEY. E, F. R. 2013. Sex differences in pain: a brief review of clinical and experimental findings. *Brittish Journal of Anaesthesia,* 111**,** 52-8.

BINGEFORS, K. & ISACSON, D. 2004. Epidemiology, co-morbidity, and impact on health-related quality of life of self-reported headache and musculoskeletal pain--a gender perspective. *Eur J Pain,* 8**,** 435-50.

BOHANNON, R. W. 1997. Comfortable and maximum walking speed of adults aged 20—79 years: reference values and determinants. *Age and Ageing,* 26**,** 15-19.

CARLOZZI, N. E., KRATZ, A. L., DOWNING, N. R., GOODNIGHT, S., MINER, J. A., MIGLIORE, N. & PAULSEN, J. S. 2015. Validity of the 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) in individuals with Huntington disease (HD). *Qual Life Res,* 24**,** 1963-71.

DE BOECK P, W. M. 2004. Explanatory Item Response Models. *In:* DE BOECK P , W. M. (ed.) *A Generalized Linear and Nonlinear Approach.* Springer-Verlag New York.

DEWITT R 2013. A Study of the Sit-up Type of Test as a Means of Measuring Strength and Endurance of the Abdominal Muscles. *Research Quarterly. American Association for Health, Physical Education and Recreation,* 15**,** 4.

GOMEZ-OLIVE, F., SCHRÖDERS, J. & ABODERIN, I. 2017. Variations in disability and quality of life with age and sex between eight lower income and middle-income countries: data from the INDEPTH WHO-SAGE collaboration. *BMJ Glob Health,* 2.

KIRCHBERGER, I., BRAITMAYER, K. & COENEN, M. 2014. Feasibility and psychometric properties of the German 12-item WHO Disability Assessment Schedule (WHODAS 2.0) in a population-based sample of patients with myocardial infarction from the MONICA/KORA myocardial infarction registry. *Population Health Metrics,* 12**,** 13.

KUTLAY, Ş., KÜÇÜKDEVECI, A. A., ELHAN, A. H., ÖZTUNA, D., KOÇ, N. & TENNANT, A. 2011. Validation of the World Health Organization disability assessment schedule II (WHODAS-II) in patients with osteoarthritis. *Rheumatology International,* 31**,** 339-346.

LUCIANO, J. V., AYUSO-MATEOS, J. L., AGUADO, J., FERNANDEZ, A., SERRANO-BLANCO, A., ROCA, M. & HARO, J. M. 2010. The 12-item World Health Organization Disability Assessment Schedule II (WHO-DAS II): a nonparametric item response analysis. *BMC Med Res Methodol,* 10**,** 45.

MASSY-WESTROPP, N., GILL, T., TAYLOR, A., BOHANNON, R. & HILL, C. 2011. Hand Grip Strength: age and gender stratified normative data in a population-based study. *BMC Research Notes,* 4**,** 127.

MERRIL S, SEEMAN T, KASL S & L, B. 1997. Gender Differences in the Comparison of Self-Reported Disability and Performance Measures. *Journal of Gerontology: MEDICAL SCIENCES,* 52A**,** 8.

NOVAK, S., COLPE, L., BARKER, P. & GFROERER, J. 2010. Development of a brief mental health impairment scale using a nationally representative sample in the USA. *International_Journal_of_Methods_in_Psychiatric_Research***,** 11.

SALTYCHEV, M., BARLUND, E., MATTIE, R., MCCORMICK, Z., PALTAMAA, J. & LAIMI, K. 2017. 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. *Clin Rehabil,* 31**,** 262-272.

SWAMINATHAN H AND ROGERS HJ 1990. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement* 27**,** 361-370.

USTUN, T. B., CHATTERJI, S., KOSTANJSEK, N., REHM, J., KENNEDY, C., EPPING-JORDAN, J., SAXENA, S., VON KORFF, M., PULL, C. & PROJECT, W. N. J. 2010. Developing the World Health Organization Disability Assessment Schedule 2.0. *Bull World Health Organ,* 88**,** 815-23.

WORLD HEALTH ORGANISATION 2010. *Measuring Health and Disability: Manual for WHO Disability Assessment Schedule (WHODAS 2.0)*, WHO Press, World Health Organization.

Figure 1. Item characteristics curves of WHODAS 2.0 items 1 to 12

Estimates for men are presented in dash lines, estimates for women (or the entire sample) – in solid lines. Y-axis presenting the probability of endorsing the item, X-axis presenting the ability level

S1 = standing S2 = household responsibilities S3 = learning S4 = joining community S5 = emotional affection, S6 = concentrating, S7 = walking, S8 = washing, S9 = dressing, S10 = dealing with strangers, S11 = maintaining a friendship, S12 = work

IV

MINIMAL CLINICALLY IMPORTANT DIFFERENCE AND MINIMAL DETECTABLE CHANGE OF THE WORLD HEALTH ORGANIZATION DISABILITY ASSESSMENT SCHEDULE 2.0 (WHODAS 2.0) AMONGST PATIENTS WITH CHRONIC MUSCULOSKELETAL PAIN

by

Niina Katajapuu, Ari Heinonen & Mikhail Saltychev, 2020

Clinical Rehabilitation vol 34, 1506-1511

DOI 10.1177/0269215520942573

Original Article

CLINICAL REHABILITATION

Clinical Rehabilitation
2020, Vol. 34(12) 1506–1511
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0269215520942573
journals.sagepub.com/home/cre
SAGE

Minimal clinically important difference and minimal detectable change of the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) amongst patients with chronic musculoskeletal pain

Niina Katajapuu[1,2], Ari Heinonen[1] and Mikhail Saltychev[3]

Abstract

Objectives: The aim of this study is to estimate a minimal clinically important difference (MCID) and a minimal detectable change (MDC) of the 12-item WHODAS 2.0 amongst patients with chronic musculoskeletal pain.

Design: Cross-sectional cohort study.

Setting: Outpatient Physical and Rehabilitation Medicine clinic.

Subjects: A total of 1988 consecutive patients with musculoskeletal pain.

Interventions: A distribution-based approach was employed to estimate a minimal clinically important difference, a minimal detectable change, and a minimal detectable percent change (MDC%).

Results: The mean age of the patients was 48 years, and 65% were women. The average intensity of pain was 6,3 (2.0) points (0–10 numeric rating scale) and the mean WHODAS 2.0 total score was 13 (9) points out of 48. The minimal clinically important difference ranged between 3.1 and 4.7 points. The minimal detectable change was 8.6 points and minimal detectable % change was unacceptably high 66%.

Conclusions: Amongst patients with chronic musculoskeletal pain, the 12-item WHODAS 2.0 demonstrated a high minimal detectable change of almost nine points. As the minimal detectable change exceeded the level of minimal clinically important difference, nine points were considered to be the amount of change perceived by a respondent as clinically significant.

Keywords

Whodas, minimal clinically important difference, minimal detectable change, musculoskeletal pain

Received: 9 January 2020; accepted: 25 June 2020

[1]Faculty of Sports and Health Sciences, University of Jyväskylä, Jyväskylä, Finland
[2]Faculty of Health and Wellbeing, Turku University of Applied Sciences, Turku, Finland
[3]Department of Physical and Rehabilitation Medicine, Turku University Hospital and University of Turku, Turku, Finland

Corresponding author:
Niina Katajapuu, Turku University of Applied Sciences, Joukahaisenkatu 3, Turku, 20520, Finland.
Email: niina.katajapuu@turkuamk.fi

Introduction

The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) is a generic tool to assess functioning in diverse situations.[1–3] While the WHODAS 2.0 has widely been used in clinical practice and research, the interpretation of results, obtained from the WHODAS 2.0 responses, have not been well defined.[4,5]

The interpretation of test results relies heavily on such characteristics as minimal clinically important difference ("MCID") and minimal detectable change ("MDC"). While changes in test score may be statistically significant, they are not necessarily perceived by patients as clinically significant. This is especially true when the results are obtained from a large sample – a very small difference may become statistically significant while its practical importance perceived by patients is negligible. Minimal clinically important difference describes the smallest amount of change or difference that might be considered important by patients or clinicians.[6] There are two common method to calculate a minimal clinically important difference: an anchor-based and a distribution-based. There is no general agreement on which method is preferable. Probably, they both have their pros and cons in different particular situations. The anchor can be either an objective or subjective measure (e.g. question about mild improvement noticed by a patient or a clinician). An anchor-based method reflects the patient's of clinician's point of view. In turn, a distribution-based method is based explicitly on the statistical variability of obtained scores. A minimal detectable change is the smallest amount of change or difference that is not the result of measurement error.[7] In an clinically ideal world, the minimal clinically important difference must exceed the level of minimal detectable change. If minimal clinically important difference is less than minimal detectable change, then the observed result below the level of minimal detectable change may be caused by chance and not by the true difference in scores even if the result exceeds the level of minimal clinically important difference.[8]

Thus, without knowledge on minimal clinically important difference and minimal detectable change, the clinical meaning of the WHODAS 2.0 total score estimates remains unclear. The minimal clinically important difference of the 12-item WHODAS 2.0 has been established by a single study amongst patients with anxiety and stress disorders.[9] In that study, an anchor-based method has been used, and the minimal clinically important difference has been estimated around three points for a less strict model and six to seven points for a stricter model. The minimal detectable change of the WHODAS 2.0 has also recently been reported by a single study amongst institutionalized ambulatory older adults as 10 points.[10] So far, there have not been reports on the minimal clinically important difference or minimal detectable change of the WHODAS 2.0 amongst patients with musculoskeletal health conditions. Due to the WHODAS 2.0 score's multidimensionality and, thus, potentially high level of minimal detectable change, the trustworthiness of the WHODAS 2.0 total score has been questioned.[2,11]

The objective of this study was to estimate the minimal clinically important difference and minimal detectable change of 12-item WHODAS 2.0 amongst people with chronic musculoskeletal pain.

Methods

Data for this study were derived from the Turku ICF Study (T54/2012) approved by the ethics committee of Hospital District of Southwest Finland (ETMK 60/180/2012). Participants provided their written informed consent for participation. This was a cross-sectional study amongst 3150 consecutive patients who were seen in an outpatient Physical and Rehabilitation Medicine clinic of university hospital between April 2014 and February 2017. The survey was sent to the patients and filled up before a physician appointment. The survey included 12-item WHODAS 2.0 questionnaire and questions on demographics, pain intensity, and perceived general health.

Self-administered 12-item WHODAS 2.0

The self-administered 12-item WHODAS 2.0 contains 12 items covering the most common limitations of functioning appearing in general population

(Appendix A). The questionnaire covers limitations during the last 30 days. A Likert-like scale is used to define the severity of limitation with 0 denoting "no limitation" and 4 denoting "extreme limitation or inability to function." The total score is the sum of all 12 items where a score of 48 points represents the worst possible restriction.[1]

Independent variables

Age was defined in full years at the time of visiting the clinic. *Pain intensity* was assessed using a 11-point numeric rating scale with 0 denoting "no pain" and 10 denoting "worst possible pain." *Educational level* was dichotomized "further education" (equivalent "further education or higher" in UK) versus "no further education" (equivalent of "primary and secondary education" in UK). *Body mass index* was calculated as a body mass divided by a squared body height (kg/m^2). *Perceived general health* status was assessed on a 4-point scale where 0 indicated best possible and 3 worst possible health. Main diagnoses were defined using the International Classification of Diseases, 10th edition

Statistical analysis

The results were reported as means, standard deviations, and standard errors, medians, ranges, and interquartile ranges when appropriate. The internal consistency was assessed by Cronbach's alpha considering $\alpha \geqslant 0.9$ excellent, $0.8 \leqslant \alpha < 0.9$ good, $0.7 \leqslant \alpha < 0.8$ acceptable, $0.6 \leqslant \alpha < 0.7$ questionable, $0.5 \leqslant \alpha < 0.6$ poor, and $\alpha < 0.5$ unacceptable.

To describe the variability between an individual's observed score and the true score, standard error of measurement (SEM) was calculated as SEM = SD × $\sqrt{}(1 - r_{xx})$ where r_{xx} is reliability coefficient of the test – in this case, Cronbach's alpha.[12] Since the data were cross-sectional and no patients' opinion on perceived change in functioning was available as an anchor, a distribution-based approach was employed to estimate minimal clinically important difference for the WHODAS 2.0. Three different formulas were used for the task:[13–18]

1) Minimal clinically important difference = standard error of measurement
2) Minimal clinically important difference = 0.5 × standard deviation
3) Minimal clinically important difference = 0.33 × standard deviation

The minimal detectable change was calculated as 1.96 × standard error of measurement × $\sqrt{}2$. The minimal detectable change was also expressed as a percentage ("MDC%") – an estimate that is independent of the units of measurement. Representing the relative amount of random measurement error, the minimal detectable % change was calculated as (minimal detectable change /observed mean WHODAS total score) × 100. The minimal detectable % change <30% was considered acceptable and <10% excellent[19,20]

All the analyses were conducted using Stata/IC Statistical Software: Release 15. College Station (StataCorp LP, TX, USA).

Results

Of 3150 patients visiting the clinic, 1988 (63%) participated the study. The patients were 47.6 (6.3) years old and 1297 (65%) were women. The average intensity of pain was 6.3 (2.0) points on a numeric rating scale. The general health median was 1 (range 0 to 4, IQR 1 to 2) (Table 1). The majority of the patients were referred to the clinic due to non-specific chronic pain in their low back, neck, extremities, or soft tissue in general. Due to the national guidelines, patients with rheumatoid arthritis, severe osteoarthritis, or fractures were referred to other specialized clinics. Thus, only one patient had a main diagnosis of rheumatoid arthritis, 0.3% had diagnoses of traumas, and 2% had diagnoses of primary osteoarthritis. Most of the patients (*n* = 1746, 88%) had a main diagnosis "M" – "Diseases of the musculoskeletal system and connective tissue." The most frequent single diagnoses were "M54 Dorsalgia" (*n* = 781, 39%) and "M79 Other soft tissue disorders" (*n* = 202, 10%). The patients' characteristics are presented in Table 1.

The distribution of WHODAS 2.0 total score was abnormal with shift towards mild disability

Table 1. Demographic characteristics and the WHODAS 2.0 total score.

Variable	Total
Age (mean and standard deviation), years	47.6 (6.3)
Body mass index (mean and standard deviation), kg/cm^2	27.4 (5.7)
Pain (mean and standard deviation), points	6.3 (2.0)
Educational level (absolute proportions and percentage)	
No further education	1258 (67%)
Further education or higher	609 (33%)
WHODAS 2.0 (mean and standard deviation), points	13.1 (9.4)
WHODAS 2.0 (median, range, and interquartile range [IQR]), points	12 (0 to 48, IQR 6 to 19)

WHODAS: World health Organization Disability Assessment Schedule.

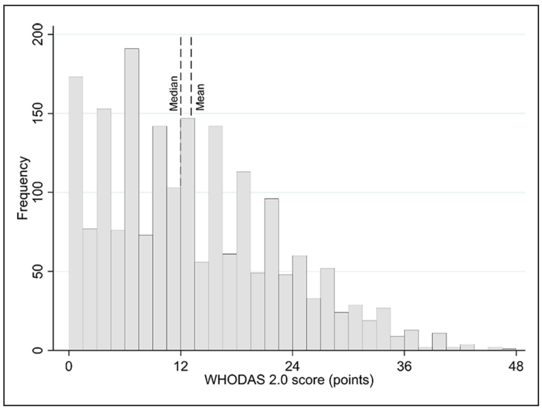


Figure 1. Histogram of the WHODAS 2.0 total score distribution.
WHODAS: World Health Organization Disability Assessment Schedule.

levels (Figure 1). However, the median and mean estimates were alike and thus, the distribution was considered close to normal enough to proceed with calculations based on mean and standard deviation. The Cronbach's alpha was good 0.89. The mean WHODAS 2.0 total score was 13.1 (9.4) and median 12 (Inter quartile range 6–19, range 0–48) points. Based on three different calculation formulas, the minimal clinically important difference estimates for the WHODAS 2.0 were 3.10 (calculated as standard error of measurement), 3.09 (calculated as one third of standard deviation), to 4.68 (calculated as half of standard deviation) points. The minimal detectable change was 8.6 points exceeding the level of minimal clinically important difference and minimal detectable percentage change was unacceptably high 66%.

Discussion

Amongst almost 2000 patients with chronic musculoskeletal pain, the 12-item WHODAS 2.0 total score demonstrated a minimal clinically important difference of three to five points with minimal detectable change of up to 9 points exceeding a minimal clinically important difference almost twice. The minimal detectable % change showed that almost 70% change in WHODAS 2.0 total score should be expected before patients or clinicians might detect the change clinically.

While the large study sample advocates for the trustworthiness of the findings, the generalization of the results may be compromised by the study's cross-sectional design. Indeed, there were not longitudinal data to re-check the estimates by using an anchor-based approach with patients' real responses on the changed clinical situation. It has to be kept in mind that minimal clinically important difference and minimal detectable change are always statistical approximations, which could be different in real-world circumstances. The WHODAS 2.0 scores were distributed abnormally in the studied sample with most of the patients perceived only mild limitations of functioning. Therefore, caution is needed when generalizing the results amongst populations with more severe limitations.

Consistent with the results of this study, a study by Silva et al. has recently set the minimal detectable change of 12-item WHODAS in institutionalized elderly to 9.6 points.[10] Respectively, the size of minimal clinically important difference seen in the present study was similar to the estimates reported previously by a study amongst patients with anxiety and stress disorders.[9] The results indirectly support previous reports on the potential unreliability of WHODAS 2.0 total score due to multidimensionality and a significant floor effect. A recent review suggested that 12-item WHODAS 2.0 is a multidimensional scale and it might be more useful when used to create a functioning profile than when providing a single total score.[2] A substantial floor effect of the 12-item WHODAS 2.0 has been seen in two studies.[21,22] All these previous findings may explain the high estimates of minimal clinically important difference and minimal detectable change seen in the present study.

Further research in different populations is recommended. To confirm the results by using an anchor-based approach, longitudinal design is needed. The WHODAS 2.0 can be scored using a simple addition of individual items' scores (used in this study) or a more complex scheme taking into account the weights of different domains included into the WHODAS 2.0. Using that second scheme might affect the observed estimates and this possibility could be investigated by further research.

Clinical Messages

- Amongst patients with chronic musculoskeletal pain, the 12-item WHODAS 2.0 demonstrated a high minimal detectable change of almost nine points.
- As the minimal detectable change exceeded the level of minimal clinically important difference, nine points were considered to be the amount of change perceived by a respondent as being clinically significant.

Acknowledgements

None to declare

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Niina Katajapuu https://orcid.org/0000-0002-7416-8928
Mikhail Saltychev https://orcid.org/0000-0003-1269-4743

References

1. Üstün TB, Kostanjsek N and Chatterji S. *Measuring Health and Disability: Manual for WHO Disability Assessment Schedule (WHODAS 2.0)*. Geneva: World Health Organization, 2010.
2. Saltychev M, Katajapuu N, Barlund E, et al. 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. *Disabil Rehabil* 2019: 1–6.
3. Carlozzi NE, Kratz AL, Downing NR, et al. Validity of the 12-item World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) in individuals with Huntington disease (HD). *Qual Life Res* 2015; 24: 1963–1971.
4. Von Korff M, Katon WJ, Lin EHB, et al. Functional outcomes of multi-condition collaborative care and successful ageing: results of randomised trial. *BMJ* 2011; 343: d6612–d6612.
5. Ferraz DD, Trippo KV, Duarte GP, et al. The effects of functional training, bicycle exercise, and exergaming on walking capacity of elderly patients with Parkinson disease: a pilot randomized controlled single-blinded trial. *Arch Phys Med Rehabil* 2018; 99: 826–833.
6. Guyatt G, Walter S and Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987; 40(2): 171–178.
7. Beaton DE, Bombardier C, Katz JN, et al. A taxonomy for responsiveness. *J Clin Epidem* 2001; 54: 1204–1217.
8. de Vet HCW and Terwee CB. The minimal detectable change should not replace the minimal important difference. *J Clin Epidem* 2010; 63(7): 804–805.
9. Axelsson E, Lindsater E, Ljotsson B, et al. The 12-item self-report World Health Organization Disability Assessment Schedule (WHODAS) 2.0 administered via the Internet to individuals with anxiety and stress disorders: a psychometric investigation based on data from two clinical trials. *JMIR Ment Health* 2017; 4: e58.
10. Silva AG, Cerqueira M, Raquel Santos A, et al. 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. *Disabil Rehabil* 2019; 41(3): 366–373.

11. Saltychev M, Mattie R, McCormick Z, et al. Confirmatory factor analysis of 12-Item World Health Organization Disability Assessment Schedule in patients with musculoskeletal pain conditions. *Clin Rehabil* 2017; 31(5): 702–709.
12. Kohn CG, Sidovar MF, Kaur K, et al. Estimating a minimal clinically important difference for the EuroQol 5-Dimension health status index in persons with multiple sclerosis. *Health Qual Life Outcomes* 2014; 12: 66.
13. Gilbert C, Brown MCJ, Cappelleri JC, et al. Estimating a minimally important difference in pulmonary arterial hypertension following treatment with sildenafil. *Chest* 2009; 135(1): 137–142.
14. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002; 77(4): 371–383.
15. Le QA, Doctor JN, Zoellner LA, et al. Minimal clinically important differences for the EQ-5D and QWB-SA in Post-traumatic Stress Disorder (PTSD): results from a Doubly Randomized Preference Trial (DRPT). *Health Qual Life Outcomes* 2013; 11: 59.
16. Pickard AS, Neary MP and Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health Qual Life Outcomes* 2007; 5: 70.
17. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008; 61(2): 102–109.
18. Shikiar R, Harding G, Leahy M, et al. Minimal important difference (MID) of the Dermatology Life Quality Index (DLQI): results from patients with chronic idiopathic urticaria. *Health Qual Life Outcomes* 2005; 3: 36.
19. Smidt N, van der Windt DA, Assendelft WJ, et al. Interobserver reproducibility of the assessment of severity of complaints, grip strength, and pressure pain threshold in patients with lateral epicondylitis. *Arch Phys Med Rehabil* 2002; 83(8): 1145–1150.
20. Chen CH, Lin SF, Yu WH, et al. Comparison of the test-retest reliability of the balance computerized adaptive test and a computerized posturography instrument in patients with stroke. *Arch Phys Med Rehabil* 2014; 95(8): 1477–1483.
21. Katajapuu N, Laimi K, Heinonen A, et al. Floor and ceiling effects of the World Health Organization Disability Assessment Schedule 2.0 among patients with chronic musculoskeletal pain. *Int J Rehabil Res* 2019; 42(2): 190–192.
22. Gaskin CJ, Lambert SD, Bowe SJ, et al. Why sample selection matters in exploratory factor analysis: implications for the 12-item World Health Organization Disability Assessment Schedule 2.0. *BMC Med Res Methodol* 2017; 17: 40.

Appendix A. The 12-item WHODAS 2.0

In the past 30 days, how much difficulty did you have in:

1. Standing for long periods such as 30 minutes?
2. Taking care of your household responsibilities?
3. Learning a new task, for example, learning how to get to a new place?
4. How much of a problem did you have in joining in community activities (e.g. festivities, religious or other activities) in the same way as anyone else can?
5. How much have you been emotionally affected by your health problems?
6. Concentrating on doing something for 10 minutes?
7. Walking a long distance such as a kilometer (or equivalent)?
8. Washing your whole body?
9. Getting dressed?
10. Dealing with people you do not know?
11. Maintaining a friendship?
12. Your day-to-day work?

(0) None; (1) Mild; (2) Moderate; (3) Severe; (4) Extreme/Cannot do