Cross-diagnostic scale-banking using rasch analysis: developing a common reference metric for generic and health condition-specific scales in people with rheumatoid arthritis and stroke

Birgit Prodinger, A. Küçükdeveci, S. Kutlay, A. Elhan, S. Kreiner, A. Tennant

Angaben zur Veröffentlichung / Publication details:

Prodinger, Birgit, A. Küçükdeveci, S. Kutlay, A. Elhan, S. Kreiner, and A. Tennant. 2020. "Cross-diagnostic scale-banking using rasch analysis: developing a common reference metric for generic and health condition-specific scales in people with rheumatoid arthritis and stroke." *Journal of Rehabilitation Medicine* 52 (10): 1–10. https://doi.org/10.2340/16501977-2736.

J Rehabil Med 2020; 52: jrm00107

ORIGINAL REPORT

CROSS-DIAGNOSTIC SCALE-BANKING USING RASCH ANALYSIS: DEVELOPING A COMMON REFERENCE METRIC FOR GENERIC AND HEALTH CONDITION-SPECIFIC SCALES IN PEOPLE WITH RHEUMATOID ARTHRITIS AND STROKE

Birgit PRODINGER, PhD[1,2,3], Ayşe A. KÜÇÜKDEVECI, MD[4], Sehim KUTLAY, MD[4], Atilla H. ELHAN, PhD[5], Svend KREINER, PhD[6] and Alan TENNANT, PhD[2,3,7]

From the [1]Faculty of Applied Health and Social Sciences, Technical University of Applied Sciences Rosenheim, Rosenheim, Germany, [2]Swiss Paraplegic Research, Nottwil, [3]Department of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland, [4]Department of Physical Medicine and Rehabilitation, Faculty of Medicine, Ankara University, [5]Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkey, [6]Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark and [7]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK

Objectives: **To develop a common reference metric of functioning, incorporating generic and health condition-specific disability instruments, and to test whether this reference metric is invariant across 2 health conditions.**

Design: **Psychometric study using secondary data analysis. Firstly, the International Classification of Functioning, Disability and Health (ICF) Linking Rules were used to examine the concept equivalence between the World Health Organization Disability Assessment Schedule (WHODAS 2.0), Health Assessment Questionnaire (HAQ) and Functional Independence Measure (FIM™). Secondly, a scale-bank was developed using a reference metric approach to test-equating, based on the Rasch measurement model.**

Participants: **Secondary analysis was performed on data from 487 people; 61.4% with rheumatoid arthritis and 38.6% with stroke.**

Results: **Three sub-domains of the WHODAS 2.0 and all items of the HAQ and FIM™ motor mapped on to the ICF chapters d4 Mobility, d5 Self-care and d6 Domestic life. Test-equating of these scales resulted in good model fit, indicating that a scale bank and associated reference metric across these 3 instruments could be created.**

Conclusion: **This study provides a transformation table to enable direct comparisons among instruments measuring physical functioning commonly used in rheumatoid arthritis (HAQ) and stroke (FIM™ motor scale), as well as in people with disability in general (WHODAS 2.0).**

Key words: psychometrics; outcome assessment; stroke; rheumatoid arthritis; World Health Organization Disability Assessment Schedule; WHODAS 2.0; Functional Independence Measure; FIM™; Health Assessment Questionnaire; HAQ.

Accepted Aug 18, 2020; Epub ahead of print Sep 8, 2020

J Rehabil Med 2020; 52: jrm00107

Correspondence address: Birgit Prodinger, Technical University of Applied Sciences Rosenheim Hochschulstr. 1, 83024 Rosenheim, Germany. E-mail: birgit.prodinger@th-rosenheim.de

LAY ABSTRACT

Functioning is what matters most to people with chronic health conditions, such as stroke or rheumatoid arthritis. While medical signs and symptoms related to these health conditions may vary widely, research has shown that people may experience similar problems with functioning. Therefore, being able to monitor and compare functioning over time is essential for the planning and allocation of rehabilitation. This study provides evidence that a common measure can be created, based on a single general disability instrument and 2 health condition-specific instruments. For clinical practice this implies that standardized reporting of functioning can be achieved based on a common measure, while data collection can continue using the commonly used and established instruments.

Functioning, as the third health indicator in health systems, complements information about mortality and morbidity by providing information about how a health condition plays out in everyday life (1). Functioning includes information about what a person does in everyday life, including moving around, getting dressed, doing housework or participating in paid work, as well as the interaction of these activities with the health condition, impairments in body structures and functions, and with contextual factors. A detriment in any domain of functioning refers to disability (2). The number of people living with disability worldwide is increasing steadily (3). For various chronic health conditions, including stroke and rheumatoid arthritis (RA), although indicators of mortality and morbidity are declining, the number of people who experience long-term disability after having been diagnosed with such a health condition is increasing (4, 5).

Rehabilitation is a strategy aimed at optimizing functioning (6). As such, it is essential to monitor functioning and disability, as well as to set targeted interventions at the individual and population level. Nevertheless, a lack of data on functioning has been continuously reported (7). Functioning information is predominantly collected with a focus on single health conditions. While this may be justified and necessary for certain purposes, it has been shown that people with various disorders, including stroke, multiple sclerosis,

doi: 10.2340/16501977-2736

and RA experience similar functioning problems in their everyday life despite different underlying health conditions (8). Consequently, in order to compare functioning across diverse conditions, information is needed that is invariant across those conditions. Invariance implies that, at the same level of functioning, an instrument measuring functioning has the same meaning and yields a comparable score across relevant groups.

At least 2 approaches can be utilized for documenting and reporting functioning information that is invariant across health conditions. First, generic disability instruments can be used that have been shown to be both reliable and valid across the relevant health conditions. Secondly, transformation tables can be established, that enable the comparison of disability scores using different instruments across health conditions. Regarding the first approach, the World Health Organization Disability Assessment Schedule (WHODAS 2.0) is a generic disability instrument, which has been translated into various languages and its psychometric properties have been tested in various health conditions (9). However, no study has been conducted to date that has examined whether the WHODAS 2.0 is invariant across different health condition groups, such as musculoskeletal and neurological disorders. Regarding the second approach, previous research has established the principles of how to develop a transformation table allowing the reporting of scores of different instruments on a reference metric (10). To our knowledge, to date, no study has examined whether a reference metric can be established across multiple scales from different health condition groups.

OBJECTIVE

The objective of this study was to create a reference metric underlying instruments commonly used along the continuum of care to measure functioning domains in people with various chronic health conditions. More specifically, the aims were:

- to develop a reference metric of functioning, incorporating generic and health condition-specific disability instruments;
- to test whether this metric is invariant across a neurological (stroke) and a musculoskeletal (RA) health condition.

A "reference metric" is defined here as one upon which 3 or more instruments are calibrated, whereas a "common metric" is a co-calibration of 2 instruments.

METHODS

A psychometric study was conducted using secondary analysis of data collected previously. A common item, non-equivalent person design was deployed in an innovative manner by using the total scores from scales as partial credit items in order to equate tests (11). Test-equating applications have a long tradition in education and psychology (12), whereas their application in health was rare until recently, where, for example, one study linked 6 sleep disorder scales based on an ordinal reference metric using the Leunbach's model (13). The current study uses Andrich's RUMM2030 (14) to equate 3 instruments widely applied in health outcome studies, to create an interval scale reference metric, upon which each of the 3 scales are calibrated via the metric, and to test that the reference metric is invariant across age, sex and different health conditions.

Sample

The RA set included data for 299 outpatients with RA who responded to questions in the WHODAS 2.0 and the Health Assessment Questionnaire (HAQ) for a previous methodological outcome measurement study (15). The stroke set included data for 188 community-dwelling patients living with stroke who completed the WHODAS 2.0 and the Functional Independence Measure (FIM™) for a previous validation study (16). For all 3 instruments the validated Turkish versions were administered (17, 18). Both studies were performed at the Department of Physical Medicine and Rehabilitation, Ankara University Medical Faculty and the ethical approval was given by the Research Ethics Committee of Medical Faculty, Ankara University, study number 127-3559 (for the study related to the RA set) and 136-3990 (for the study related to the stroke set).

Instruments

The World Health Organization Disability Assessment Schedule (WHODAS 2.0) is a generic disability instrument. The complete version consists of 36 items on functioning and disability within 6 domains: understanding and communicating (6 items), getting around (5 items), self-care (4 items), getting along with others (5 items), life activities (8 items), and participation in society (8 items). Four of the latter relate to school or work situations, and can be omitted if not relevant. Items are scored on a 5-point scale, ranging from 1 = none to 5 = extreme/cannot do. Six domain scores and a total score are available for the evaluation of dimensions of disability and health status; higher scores reflect greater disability (19). WHODAS 2.0 has been tested and used in more than 16 countries, mainly among adults 18 years of age or above. Both classical and modern psychometric analyses have been used to support the validity of the instrument in RA and stroke populations (9). We included only those domains of the WHODAS 2.0 in our analyses that revealed conceptual equivalence with the other instruments included in this study. It is noteworthy that the few previous Rasch analyses conducted of the WHODAS 2.0 in specific health conditions focused on generating a score on the full scale rather than a score at the domain level (20, 21).

With respect to health condition-specific measures, data on the HAQ was collected in patients with RA, and data on the FIM™ in patients with stroke. The HAQ was developed to be used across various rheumatic conditions (22) and has been described as a valid, reliable and responsive measure in the RA population (23). The HAQ consists of 20 items divided into 8 domains: Dressing & Grooming, Arising, Eating, Walking, Hygiene, Reach, Grip, and Activities. All items are rated on a 4-point scale (0 = without any difficulty, 3 = unable to do). The highest score reported by the patient for any question within each domain determines the score for that domain. Subsequently, the mean score of the 8 domains is calculated as the HAQ score in a range of 0–3. In this study, the HAQ was scored without the score adjustment for assistive

devices and help. Since the other included PROMs reflect a performance perspective, whereas adjusting HAQ scores attempts a capacity perspective, i.e. trying to ascertain what level of problem the individual would have had without using assistive devices or help, we refrained from the score adjustment.

The Functional Independence Measure (FIM™) motor scale is a widely used generic assessment tool, which can be used as an outcome measure for the functional status and burden of care in rehabilitation patients (24). The FIM™ also includes a cognitive scale, which was not used in this study. The FIM™ motor scale consists of 13 items, which can be grouped into 3 sub-scales: self-care; sphincter control; and transfer and mobility (25). A 7-level scoring system is used to rate independence in each item, where 1 = complete dependence and 7 = complete independence. Thus, the total score ranges from 13 to 91, where higher scores indicate higher functional independence. Studies of the psychometric quality of the FIM™ have shown that it has a high overall internal consistency, adequate discriminative capabilities for rehabilitation patients and some responsiveness, construct validity, and good inter-rater reliability (26, 27). Furthermore, previous Rasch analyses of the FIM™ have shown that there are local dependencies amongst items, which can be absorbed by replacing the dependent items with testlet scores (28).

Data collection

WHODAS 2.0 was collected in both clinical populations, whereas FIM™ was collected only in people with stroke, and the HAQ only in people with RA.

Data analysis

To establish comparability of existing scales, 2 aspects are important (10). First, to examine the conceptual equivalence of the existing instruments, they were linked to a universal reference framework. The International Classification of Functioning, Disability and Health (ICF) was used in this study, which is the recommended standard set out by the World Health Organization to describe health and disability of individuals and populations, providing an internationally agreed language and structure (2). The ICF Linking Rules, an established method to link existing instruments to the ICF, were applied (25). The current study accessed existing linkings from the ICF Research Branch (www.icf-research-branch.org) in which the first author was involved and which were performed accordingly. The results of the ICF Linking provide evidence for the conceptual equivalence of the identified instruments, which is fundamental for scale equating (10). For this study, we considered the items or sub-sets of items contained in the identified instruments as conceptually equivalent if they were linked to the same ICF chapter.

Secondly, to achieve score equivalence between the scores of the identified instruments, test-equating was undertaken within the Rasch measurement model framework using RUMM2030 (14). Data came from one study population with responses to FIM™ and WHODAS 2.0 and another population with responses to HAQ and WHODAS 2.0. For this reason, the data generated 2 sets of ordinal level raw scores that cannot be compared. However, under the Rasch model, the 2 sets of raw scores can be transformed into interval scaled estimates of person parameters that define the basis of a reference metric where scores from the different populations become comparable.

To establish score equivalence, 2 aspects of the study design are important: first, the WHODAS 2.0 was collected in both the RA and stroke population, and thus served as the common scale to link between the 2 data-sets. The scoring of the FIM™ motor scale was reversed so that a low score indicated no problems/high independency and a high score extreme problems/high dependency in all scales. Secondly, during test equating in RUMM2030, the total scores of the scales are equated such that the scale becomes the items (11). The location of a scale is thus the mean of the threshold locations, just as in ordinary partial credit items, except there will usually be more thresholds.

During analysis, the scales (items) are subjected to the usual Rasch analysis procedures, to test whether the data deviates from the Rasch model's assumptions of item-fit, invariance and unidimensionality. χ^2 tests and residuals are used to assess the fit of test scores (items) to the Rasch model. Due to the structural missing data design, only the WHODAS 2.0 was administered to all persons. For this reason, pairwise calibration of the WHODAS 2.0 with the HAQ and with the FIM™ motor scale was conducted before all 3 scales were equated. The pairwise analyses included a Conditional Test of Fit (CTF) of the test scores to the Rasch model to ascertain that the 2 test scores (items) measured the same latent trait. The Benjamini-Hochberg procedure was applied to adjust for multiple testing (29).

Invariance requires that 2 persons with the same trait level yet with different personal or health condition characteristics, such as male and female or condition, have the same probability of achieving a given score on the item. Under the joint model for the 3 instruments, invariance implies that there is no differential item functioning (DIF) (30) relative to age, sex, and health condition tested, in this case, by an analysis of variance (ANOVA) of the residuals.

Local response independence is an important assumption of the Rasch model (31). Items may be locally dependent because of response dependence or because of multidimensionality. For this reason, the analysis by the joint model calculated residual correlations between items (instruments in this case) and tested unidimensionality by paired *t*-tests comparing of person estimates based on the WHODAS 2.0 + HAQ with person estimates based on FIM™, and by paired *t*-tests comparing person estimates by WHODAS 2.0 + FIM™ with person estimates by HAQ. During these analyses, RUMM2030 counts the number of cases where the *p*-values of the paired *t*-tests are less than or equal to 5%, and compares this number with the expected 5% of the persons.

Once evidence for score equivalence is established, the metric needs to be defined. In principle, the person parameters of the joint Rasch model could be estimated by outcomes on the separate scores, by outcomes on WHODAS 2.0 + HAQ or WHODAS 2.0 + FIM™; or by WHODAS 2.0 + HAQ + FIM™ if data on all scores had been collected for some persons. All of these estimates would posit the persons on the same logit scale, with values from minus to plus infinity. However, since many users prefer measures without negative values, it is common to change the origin and the unit of the logit scale so that the range of possible outcomes lies within an interval from zero to a reasonable upper limit. For this reason, we propose a reference metric defined by the possible outcomes of all 3 scales. To change these logits into values with which users will be more comfortable, the origin and the unit of the logit scales were changed in such a way that the values on the WHODAS 2.0 + HAQ + FIM™ raw score transformed to an interval-scaled reference metric from 0 to 100.

RESULTS

Participants

In total, the sample consisted of 487 people; 299 (61.4%) with RA and 188 (38.6%) with stroke. In the

RA sample 25.4% were male, and in the stroke sample 53.7% were male.

Conceptual equivalence

The 3 instruments were linked to the ICF. As shown in Table I, all items of the HAQ and FIM™ motor scale were linked to the ICF chapters d4 Mobility, d5 Self-care and d6 Domestic life, as were the item blocks related to Getting around, Self-care and Life activities of the WHODAS 2.0. The item D3.4 *Staying by yourself for a few days* of the WHODAS 2.0 and the item *Do chores such as vacuuming or yard work* of the HAQ were linked to d5 Self-care and d6 Domestic life respectively, rather than to a specific ICF category, since the content of these items was not further specified.

Score equivalence

Tables II–V show the results of the analyses of fit of WHODAS 2.0, HAQ and FIM™ motor scale to the joint Rasch models for all 3 scales. There are a few significant fit statistics, but significance is generally weak, with *p*-values between 0.01 and 0.05. After adjusting for multiple testing all hypotheses were accepted, except for the evidence of DIF relative to age during pairwise calibration of WHODAS 2.0 to the other 2 scales, where the adjusted *p*-values are 0.01.

Table I. International Classification of Functioning, Disability and Health (ICF) linking table

ICF Code	ICF Label	WHODAS 2.0	HAQ	FIM™ motor
d4	Mobility			
d410	Changing basic body position	D2.2 Standing up from sitting down	Stand up from a straight chair Get in and out of bed Bend down to pick up clothing from the floor Get in and out of the car	
d415	Maintaining basic body position	D2.1 Standing for long periods such as 30 min		
d420	Transferring oneself			9 Transfers bed chair wheelchair 11 Transfer tub to shower
d430	Lifting and carrying objects		Reach and get down a 5 pound object (such as a bag of sugar) from just above your head	
d435	Moving objects with lower extremities			
d440	Fine hand use			
d445	Hand and arm use		Open car doors Open jars which have been previously opened Turn faucets on and off	
d450	Walking	D2.5 Walking a long distance such as a kilometre (or equivalent)		12 Walking or using wheelchair 13 Stairs
d455	Moving around		Climb up 5 steps	
d460	Moving around in different locations	D2.3 Moving around inside your home D2.4 Getting out of your home	Walk outdoors on flat ground	
d465	Moving around using equipment			
d470	Using transportation			
d475	Driving			
d5	Self-care	D3.4 Staying by yourself for a few days		
d510	Washing oneself	D3.1 Washing your whole body	Shampoo your hair Wash and try your body Take a tub bath	2 Grooming 3 Bathing
d520	Caring for body parts			
d530	Toileting		Get on and off the toilet	6 Toileting 7 Bladder Management 8 Bowel Management 10 Transfer Toilet
d540	Dressing	D3.2 Getting dressed	Dress yourself, including tying shoelaces and doing buttons	4 Dressing upper body 5 Dressing lower body
d550	Eating	D3.3 Eating	Cut your meat	1 Feeding
d560	Drinking		Lift a full cup or glass to your mouth Open a new milk carton	
d570	Looking after one's health			
d6	Domestic life		Do chores such as vacuuming or yard work	
d610	Acquiring a place to live	D5.1 Taking care of your household responsibilities		
d620	Acquisition of goods and services		Run errands and shop	
d630	Preparing meals	D5.2 Doing most important household tasks well		
d640	Doing housework	D5.3 Getting all the household work done that you need to do		
d650	Caring for household objects			
d660	Assisting others	D5.4 Getting your household work done as quickly as needed		

WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure.

Table II. Item (scale)-fit statistics

		Item-fit statistics											
	Sample size	WHODAS				FIM™				HAQ			
Co-calibrations	*n*	Residual	χ^2	DF	*p*-value	Residual	χ^2	DF	*p*-value	Residual	χ^2	DF	*p*-value
WHODAS+FIM™	188	0.202	17.42	7	0.015	1.100	2.07	7	0.956	–	–	–	–
WHODAS+HAQ	299	–2.834	18.42	7	0.010	–	–	–	–	1.899	8.64	7	0.280
WHODAS+FIM™+HAQ	487	–1.008	10.89	7	0.143	–1.080	2.87	7	0.897	1.560	8.99	7	0.253

WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure; DF: degrees of freedom.

Since the analysis of DIF in the joint model did not provide evidence of DIF relative to age, we accepted the joint Rasch model for the 3 scales and concluded therefore that a common reference metric for WHODAS 2.0, HAQ, and FIM™ is feasible.

Given the evidence that scale equating is possible, the logit scale was transformed into a scale from 0 to 100. Fig. 1 shows a, so-called, item-map presenting the distribution of the person estimates on the logit scale together with the locations of the items. The targeting of the equated scales was good, with a person mean of –0.417, where the item mean is 0 (Fig. 1). The slight offset to the milder end of functional limitation is driven largely by the RA sample. Fig. 2 shows the ranges of reference logits that could have resulted by a total WHODAS 2.0 + HAQ + FIM™ score, and the logit values that the separate scales could have delivered.

Table VI shows how to transform raw scores from the scales into the common reference metric. A raw score equal to, respectively, 0, 77 and 154 on WHODAS 2.0 + HAQ + FIM™ transforms to 0, 42.9 and 100 on the reference metric, while raw scores on WHODAS equal to 0, 26 and 52 correspond to reference values equal to 13.0, 38.3 and 74.0. Note that in the Table IV the FIM™ motor scores were reversed back to the original scoring direction, so that a low score indicates high dependency and a high score, low dependency.

The transformation Table VI allows clinicians and researchers to exchange information collected with the 3 instruments between each other in stroke and RA populations. It is possible to determine what the raw score on one scale would equate to on another scale, and to compare differences between raw scores in a meaningful way. Consider, for instance, the following 3 patients: patient 1 has a raw score of 37 on the summed domains of the WHODAS 2.0 that transforms to a reference metric score of 43.9; patient 2 has a raw score of 14 on HAQ corresponding to a reference metric of 39.5; and patient 3 has 52 on the FIM™ motor scale and therefore a reference score of 49.1. In other words, patient 2 has fewer and patient 3 more problems with functioning than patient 1. Assume, next, that the patient's condition has improved after rehabilitation and that we want to compare the degrees of improvement. Patient 1 has a raw score of 10 on WHODAS 2.0, patient 2 has a raw score of

Table III. Test of item trait interaction and reliability measure by the Person Separation Index (PSI)

	Item trait interaction				Reliability
Co-calibrations	Residual	χ^2	DF	*p*-value	PSI
WHODAS+FIM™	0.266	6.00	14	0.97	0.948
WHODAS+HAQ	–0.518	21.71	14	0.08	0.757
WHODAS+FIM™+HAQ	–0.176	22.76	21	0.36	0.891

WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure; DF: degrees of freedom.

Table V. Tests of unidimensionality

		Paired *t*-test				Conditional test of fit		
Co-calibrations		# of significant	*n*	%	95% CI	χ^2	DF	*p*-value
WHODAS+FIM™	W vs F	6	180	3.33	0.1–6.5	31.94	82	1.00
WHODAS+HAQ	W vs H	3	282	1.06	0.0–3.6	85.94	60	0.016
WHODAS+FIM™ +HAQ	W+F vs H	4	292	1.37	0.0–3.5	n.a.	n.a.	n.a.
WHODAS+FIM™ +HAQ	W+H vs F	4	188	2.13	0.0–5.4	n.a.	n.a.	n.a

95% CI: 95% confidence interval; n.a.: not available; WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure; DF: degrees of freedom; W: WHODAS (World Health Organization Disability Assessment Schedule); F: FIM™ (Functional Independence Measure); H: HAQ (Health Assessment Questionnaire).

Table IV. Analyses of DIF and local dependence

	Sex			Age			Health condition			Residual correlations	
Co-calibrations	F	DF	*p*-value	F	DF	*p*-value	F	DF	*p*-value	WHODAS & FIM™	WHODAS & HAQ
WHODAS+FIM™	3.765	1	0.053	6.04	3	0.001	n.a.	n.a	n.a	n.a.	n.a.
WHODAS+HAQ	2.991	1	0.031	13.84	3	0.001	n.a.	n.a	n.a.	n.a.	n.a.
WHODAS+FIM™+HAQ	0.082	1	0.775	1.427	3	0.234	5.345	1	0.021	–0.098	0.098

*Residual correlations have been adjusted to remove bias as suggested by Marais (2013).
WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure; DF: degrees of freedom.

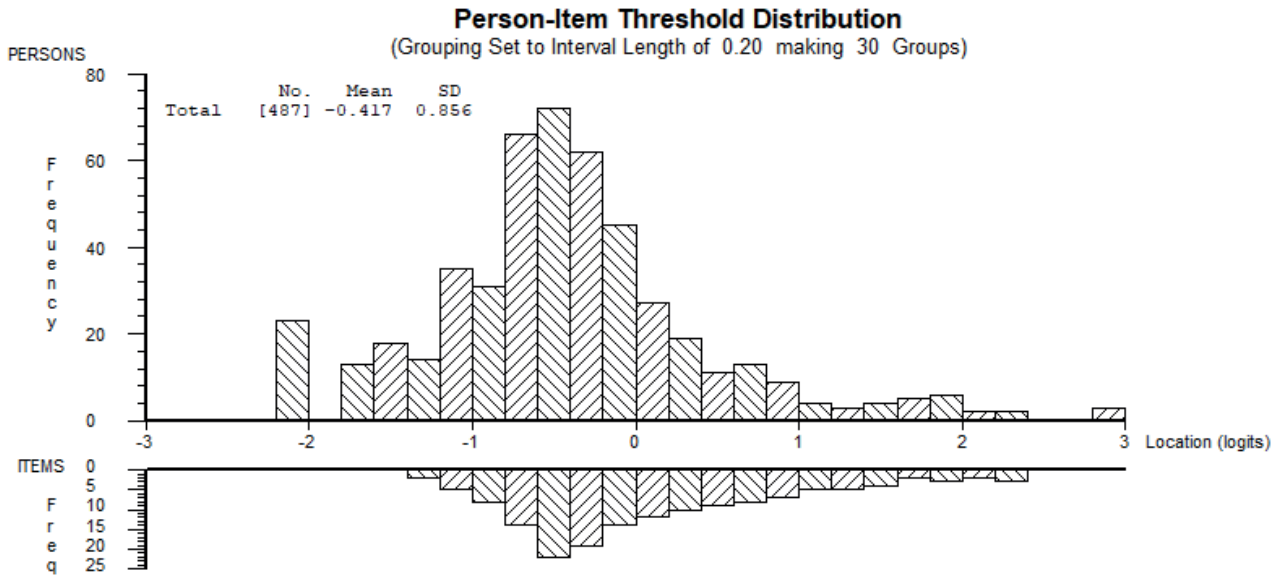

Fig. 1. Targetting of samples across the reference metric. SD: standard deviation; No.: number.

4 on HAQ and patient 3 has 75 on FIM™. Since the scores transform to, respectively, 31.4, 31.8 and 36.5 on the reference metric, we see that patients 1 and 2 are at the same level of difficulties after rehabilitation and patient 3 continues to have more difficulties. The differences in the reference scores are meaningful because the reference scale is an interval scale. These differences show that the improvement in patient 3 (12.6 on the reference metric) is more than twice the improvement in patient 2 (7.7) and marginally larger than the improvement in patient 1 (12.5).

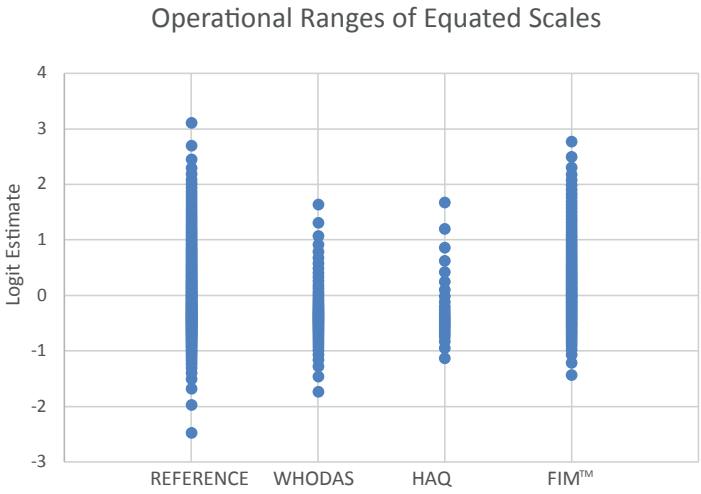

Fig. 2. Operational range of each scale in relation to the reference metric. WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure.

DISCUSSION

This study provides evidence that it is feasible to create an interval-scaled reference metric across health conditions, using existing generic and health condition-specific disability instruments, and that the reference metric was invariant across RA and stroke. The basic methods applied in this study are not new, although the integration of the ICF Linking Rules with the Rasch measurement model to establish conceptual and score equivalence between instruments has only recently been introduced (32). The ICF Linking Rules provide reference to the international standard for reporting functioning set by the World Health Organization (WHO) and endorsed by various institutes, such as the ISO Standard for Quality Management in Health Care Services (International Organization for Standardization (ISO) 9001: 2015)(33).

From one point of view, the results are similar to those from psychometric test equating of raw scores and, in particular, similar to the analysis of indirect equating (13). In this previously published study for indirect equating, the Leunbach model was utilized, where the model is the joint distribution of 2 power series distributions depending on the same person parameter. This distribution can be rewritten as the distribution of a partial credit item, and the relationship between the power series distribution and polytomous Rasch items (34–37). Thus, Leunbach's model is nothing

Table VI. Transformation table. A low score on the reference metric indicates no difficulties and a high score extreme difficulties

Reference Metric		WHODAS		HAQ			FIM™ motor	
Raw score	Ref. metric	Raw score	Ref. metric	Raw score	HAQ scoring system	Ref. metric	Raw score	Ref. metric
0	0.0	0	13.0	0	0.00	24.0	91	18.0
1	9.1	1	18.1	1	0.13	27.2	90	22.0
2	14.3	2	21.3	2	0.25	29.3	89	24.5
3	17.4	3	23.5	3	0.38	30.8	88	26.1
4	19.5	4	25.1	4	0.50	31.8	87	27.4
5	21.1	5	26.5	5	0.63	32.7	86	28.5
6	22.4	6	27.8	6	0.75	33.4	85	29.6
7	23.4	7	28.9	7	0.88	34.1	84	30.3
8	24.3	8	29.8	8	1.00	34.9	83	31.2
9	25.0	9	30.5	9	1.13	35.6	82	31.9
10	25.8	10	31.4	10	1.25	36.3	81	32.7
11	26.3	11	32.1	11	1.38	36.8	80	33.2
12	27.0	12	32.7	12	1.50	37.7	79	33.9
13	27.5	13	33.2	13	1.63	38.4	78	34.6
14	27.9	14	33.8	14	1.75	39.5	77	35.2
15	28.4	15	34.3	15	1.88	40.5	76	35.9
16	28.8	16	34.7	16	2.00	42.0	75	36.5
17	29.3	17	35.2	17	2.13	43.8	74	37.0
18	29.7	18	35.6	18	2.25	46.1	73	37.7
19	30.1	19	35.9	19	2.38	48.6	72	38.3
20	30.4	20	36.3	20	2.50	51.6	71	38.8
21	30.8	21	36.6	21	2.63	55.1	70	39.4
22	30.9	22	37.0	22	2.75	59.4	69	39.9
23	31.3	23	37.4	23	2.88	65.5	68	40.4
24	31.7	24	37.5	24	3.00	74.0	67	41.0
25	31.8	25	37.9				66	41.5
26	32.2	26	38.3				65	42.1
27	32.4	27	38.6				64	42.6
28	32.7	28	39.0				63	43.2
29	32.9	29	39.3				62	43.7
30	33.3	30	39.9				61	44.2
31	33.5	31	40.3				60	44.8
32	33.6	32	40.8				59	45.3
33	33.8	33	41.3				58	45.9
34	34.2	34	41.9				57	46.4
35	34.3	35	42.4				56	47.0
36	34.5	36	43.1				55	47.5
37	34.7	37	43.9				54	48.0
38	34.9	38	44.8				53	48.6
39	35.1	39	45.7				52	49.1
40	35.4	40	46.7				51	49.8
41	35.6	41	47.8				50	50.4
42	35.8	42	49.1				49	50.9
43	36.0	43	50.4				48	51.7
44	36.1	44	51.8				47	52.2
45	36.3	45	53.2				46	52.9
46	36.5	46	54.9				45	53.5
47	36.7	47	56.7				44	54.2
48	36.9	48	58.7				43	54.9
49	37.0	49	61.0				42	55.5
50	37.2	50	63.7				41	56.2
51	37.4	51	68.0				40	56.9
52	37.6	52	74.0				39	57.8
53	37.7						38	58.5
54	37.9						37	59.3
55	38.1						36	60.2
56	38.3						35	61.1
57	38.5						34	61.8
58	38.6						33	62.7
59	38.8						32	63.6
60	39.0						31	64.7
61	39.4						30	65.6
62	39.5						29	66.7
63	39.7						28	67.8
64	39.9						27	68.7
65	40.1						26	69.9
66	40.3						25	71.0
67	40.4						24	72.1
68	40.8						23	73.4
69	41.0						22	74.6
70	41.1						21	75.7
71	41.3						20	77.0

Table VI *cont.*

Reference Metric		WHODAS		HAQ			FIM™ motor	
Raw score	Ref. metric	Raw score	Ref. metric	Raw score	HAQ scoring system	Ref. metric	Raw score	Ref. metric
72	41.7						19	78.4
73	41.9						18	79.9
74	42.0						17	81.5
75	42.4						16	83.3
76	42.6						15	85.7
77	42.9						14	89.1
78	43.1						13	94.0
79	43.3							
80	43.6							
81	43.8							
82	44.2							
83	44.5							
84	44.7							
85	45.1							
86	45.3							
87	45.6							
88	46.0							
89	46.2							
90	46.5							
91	46.9							
92	47.0							
93	47.4							
94	47.8							
95	48.1							
96	48.3							
97	48.7							
98	49.0							
99	49.4							
100	49.7							
101	50.1							
102	50.4							
103	50.8							
104	51.2							
105	51.5							
106	51.9							
107	52.2							
108	52.6							
109	53.1							
110	53.5							
111	53.8							
112	54.2							
113	54.7							
114	55.1							
115	55.6							
116	56.0							
117	56.5							
118	56.9							
119	57.4							
120	58.0							
121	58.5							
122	59.0							
123	59.6							
124	60.1							
125	60.6							
126	61.2							
127	61.7							
128	62.4							
129	63.0							
130	63.7							
131	64.4							
132	65.1							
133	65.8							
134	66.5							
135	67.3							
136	68.2							
137	68.9							
138	69.8							
139	70.7							
140	71.6							
141	72.5							
142	73.5							
143	74.6							
144	75.7							

Table VI *cont.*

Reference Metric		WHODAS		HAQ			FIM™ motor	
Raw score	Ref. metric	Raw score	Ref. metric	Raw score	HAQ scoring system	Ref. metric	Raw score	Ref. metric
145	76.7							
146	77.8							
147	79.1							
148	80.3							
149	81.8							
150	83.4							
151	85.5							
152	88.2							
153	92.7							
154	100.0							

WHODAS: World Health Organization Disability Assessment Schedule; HAQ: Health Assessment Questionnaire; FIM™: Functional Independence Measure; Ref.: reference.

but a partial credit model with 2 items and equating at the ordinal level. However, what is innovative in the present study is that the equating was based upon an interval-scaled reference metric, derived by a joint equating model for all 3 scales, enabling a transformation from the separate scores to a joint logit scale. The logit scale is an interval scale. For this reason, the logit scale or a convenient linear transformation of the logits is the natural reference metric, on which to compare test results from the different scales.

Increasingly we are seeing studies that calibrate instruments together, usually within a single diagnosis, or sometimes with a wider focus, across, for example, a musculoskeletal group encompassing several diagnoses, or focussing on a particular symptom, such as pain (38, 39). These recent equating applications have used sample dependent item response theory (IRT) models such as the generalized partial credit model, where person estimates must be meand to provide a transformation table, given each raw score can provide, in theory, vast numbers of different estimates. The critical issue is that any transformation table presented should reflect a calibration model that delivers estimates that are independent of the distribution upon which the calibration is based. Only then, given the same frame of reference (e.g. diagnostic group(s)), can clinicians and others have confidence that those transformations apply to their own sample, involving the same frame of reference. This requires parameter separation between persons and items, which is consistent with applying the Rasch model as in the current study.

Andrich's approach, of rewriting one of the models for test equating as a partial credit model and using RUMM2030 facilities for Rasch analysis during test equating, is both recent and important for deriving the reference metric (11). Tests of invariance of WHODAS 2.0 would be a challenge for the majority of DIF tests implemented in programmes for IRT and Rasch analysis, but was not a problem for the ANOVA analysis of DIF implemented in RUMM2030. Thus, the application of these methods has significant implications for outcome research in the future. The reference metric based on these 3 instruments allows collating of data derived from any of these instruments for different purposes, such as clinical decision-making, bench marking, or meta-analyses. This approach implies that standardization of outcome measurement does not require standardization of the instruments, but rather enables the standardized reporting of functioning outcomes irrespective of the instrument used.

Study limitations

The limitations of this study are consistent with the use of existing data for secondary analysis, where no influence is possible on the initial data collection. The low percentage of males (25.4%) in the RA sample is consistent with the prevalence of RA in men. Furthermore, only data-sets with one health condition group; RA for musculoskeletal and stroke for neurological, were available that included both data on the WHODAS 2.0 and a health condition-specific instrument. The contextual factors available for DIF analysis are also constrained to only those data shared across the original studies. Note, that in the present analysis the sum-score of each scale was used for analysis. This scoring is in accordance with the traditional scoring of the WHODAS 2.0 and FIM™. The HAQ scoring is usually different as the highest score within a domain determines the score for the domain. The sum of the domain scores is then divided by the number of domains, and thus results in a score from 0 to 3. Previous research has shown fit of the 20 HAQ items to the Rasch model and emphasized the value of using the full information of all 20 items rather than the highest score of each domain (40). From the perspective of this study, it is a limitation that the HAQ scores are not directly comparable with the HAQ scorings often used in practice. Nevertheless, it has the advantage that the information from all 20 items was maintained, and if one has access to the ratings of each HAQ item, one can use the transformation table provided in this paper. Another limitation is the apparent absence of any evidence of the quality of the equating procedure. It is true that if the data fit the model (e.g. fit statistics, and graphical information), then the accuracy of every other inference to that level of fit, follows. Nevertheless, most recently a "standard error of equating" has been proposed, which shows the accuracy of equating at all score levels (13). Currently confined to 1 software package, it is hoped that this will be taken up elsewhere; for example, in the expanding R-based

Rasch procedures. Finally, the pairwise co-calibration of the WHODAS 2.0 and the HAQ resulted in a slightly less than acceptable level of non-error variance retained, but the triple calibration of all 3 scales was much more robust, and the resulting transformation tables are based on this latter analysis.

CONCLUSION

This study provides evidence and a transformation table to enable direct comparisons among instruments commonly used to measure functioning in RA (HAQ) and stroke (FIM™ motor scale), as well as in people with disability in general (WHODAS 2.0). Clinicians, public health experts and researchers are thus supported in the continuing use of their existing instruments, and their historical data collections, whilst being able, where necessary, to compare results across health conditions.

REFERENCES

1. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. Eur J Phys Rehabil Med 2017; 53: 134–138.
2. World Health Organization (WHO). International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
3. Vos T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015; 386: 743–800.
4. Woolf AD, Pfleger B. Burden of major musculoskeletal conditions. B World Health Organ 2003; 81: 646–656.
5. Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circ Res 2017; 120: 439–448.
6. Stucki G, Bickenbach J, Melvin J. Strengthening rehabilitation in health systems worldwide by integrating information on functioning in national health information systems. Am J Phys Med Rehabil 2017; 96: 677–681.
7. World Health Organization (WHO). World report on disability. Geneva: WHO; 2011.
8. Cieza A, Anczewska M, Ayuso-Mateos JL, Baker M, Bickenbach J, Chatterji S, et al. Understanding the impact of brain disorders: towards a 'horizontal epidemiology'of psychosocial difficulties and their determinants. PLOS One 2015; 10: e0136271.
9. Üstün TB. Measuring health and disability: manual for WHO Disability Assessment Schedule (WHODAS 2.0). Geneva: World Health Organization; 2009.
10. Prodinger B, Tennant A, Stucki G, Cieza A, Üstün TB. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. J Health Serv Res Pol 2016; 21: 223–228.
11. Andrich D. The polytomous Rasch model and the equating of two instruments. In: Christensen KB, Kreiner S, Mesbah M, editors. Rasch models in health. London, UK: ILSTE Ltd; 2013, p. 164–196.
12. Smith RM, Kramer GA. A comparison of two methods of test equating in the rasch model. Educ Psychol Meas 1992; 52: 835–846.
13. Adroher ND, Kreiner S, Young C, Mills R, Tennant A. Test equating sleep scales: applying the Leunbach's model. BMC Med Res Methodol 2019; 19: 141.
14. Andrich D, Sheridan B, Luo G. Rasch models for measurement: RUMM2030. Perth, Western Australia: RUMM Laboratory Pty Ltd; 2010.
15. Doğanay Erdoğan B, Elhan AH, Kaskatı OT, Öztuna D, Küçükdeveci AA, Kutlay Ş, et al. Integrating patient reported outcome measures and computerized adaptive test estimates on the same common metric: an example from the assessment of activities in rheumatoid arthritis. Int J Rheum Dis 2017; 20: 1413–1425.
16. Küçükdeveci AA, Kutlay Ş, Yıldızlar D, Öztuna D, Elhan AH, Tennant A. The reliability and validity of the World Health Organization Disability Assessment Schedule (WHODAS-II) in stroke. Disabil Rehabil 2013; 35: 214–220.
17. Küçükdeveci AA, Sahin H, Ataman S, Griffiths B, Tennant A. Issues in cross-cultural validity: example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. Arthritis Care Res 2004; 51: 14–19.
18. Küçükdeveci AA, Yavuzer G, Elhan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. Clin Rehabil 2001; 15: 311–319.
19. World Health Organization (WHO). WHO Disability Assessment Schedule 2.0 (WHODAS 2.0). Geneva: WHO; 2020. [accessed 2020 Aug 2] Available from: http: //www.who.int/classifications/icf/whodasii/en/index.html.
20. De Wolf AC, Tate RL, Lannin NA, Middleton J, Lane-Brown A, Cameron ID. The World Health Organization Disability Assessment Scale, WHODAS II: reliability and validity in the measurement of activity and participation in a spinal cord injury population. J Rehabil Med 2012; 44: 747–755.
21. Kimber M, Rehm J, Ferro MA. Measurement invariance of the WHODAS 2.0 in a population-based sample of youth. PLOS One 2015; 10: e0142385.
22. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. J Rheumatol 1982; 9: 789–793.
23. Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. J Rheumatol 2009; 36: 254–259.
24. Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles of the functional independence measure. Am J Phys Med Rehabil 1993; 72: 84–89.
25. Cieza A, Fayed N, Bickenbach J, Prodinger B. Refinements to the ICF Linking Rules to strenghten their potential for establishing comparability of health information. Disabil Rehabil 2019; 41: 574–583.
26. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. Stroke 1992; 7: 65–75.
27. Hamilton BB, Laughlin JA, Fiedler RC, Granger CV. Interrater reliability of the 7-level functional independence measure (FIM). Scand J Rehabil Med 1994; 26: 115–119.
28. Nilsson ÅL, Tennant A. Past and present issues in Rasch analysis: the Functional Independence Measure (Fim TM) revisited. J Rehabil Med 2011; 43: 884–892.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R. Statist. Soc. B, 1995; 57: 289–300.
30. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med 2000; 19: 1651–1683.
31. Marais, I. Local dependence. In: Christensen, KB, Kreiner S, Mesbah M. (eds). Rasch models in health. London: ISTE & John Wiley & Sons; 2013, p. 111–130.
32. Prodinger B, O'Connor R, Stucki G, Tennant A. Establishing score equivalence of the Functional Independence Measure Motor Scale and the Barthel Index utilizing the

International Classification of Functioning, Disability and Health and Rasch measurement theory. J Rehabil Med 2017; 49: 416–422.
33. British Standards Institution (BSI). Health care services – quality management systems. Requirements based on EN ISO 9001: 2008. London: BSI; 2012.
34. Fischer GH. The derivation of polytomous Rasch models. In: Fischer GH, Molenaar IW (editors). Rasch models – foundations, recent developments, and applications. Berlin: Springer-Verlag; 1995, p. 293–305.
35. Kreiner S, Christensen KB. Validity and objectivity in health related summated scales: analysis by graphical loglinear Rasch models. In: von Davier M, Carstensen CH (editors). Multivariate and mixture distribution Rasch models – extensions and applications. New York: Springer Verlag; 2006, p. 329–346.
36. Kreiner S. Validity and objectivity. Reflections on the role and nature of Rasch models Nordic Psychology 2007; 59: 268–298.
37. Mesbah M, Kreiner S. Rasch models for ordered polytomous items. In: Karl-Bang Christensen, Svend Kreiner, Mounir Meshah (editors). Rasch models in health. ISTE: London; 2013, p. 36.
38. Cook KF, Schalet BD, Kallen MA, Rutsohn JP, Cella D. Establishing a common metric for self-reported pain: linking BPI Pain Interference and SF-36 Bodily Pain Subscale scores to the PROMIS Pain Interference metric. Qual Life Res 2015; 24: 2305–2318.
39. Oude Vsohaar MAH, Vonkeman HE, Courvoisier D, Finckh A, Gossec L, Leung YY, et al. Towards standardized patient reported phyiscal function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Lif Res 2019; 28: 187–197.
40. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain M. Are we making the most of the Stanford Health Assessment Questionnaire? Rheumatol 1996; 35: 574–578.