AMERICAN COLLEGE
of RHEUMATOLOGY
Empowering Rheumatology Professionals

Scale Banking for Patient-Reported Outcome Measures That Measure Functioning in Rheumatoid Arthritis: A Daily Activities Metric

Birgit Prodinger,[1] (ID) Michaela Coenen,[2] Alison Hammond,[3] Ayşe A. Küçükdeveci,[4] and Alan Tennant[5]

Objective. Functioning is an important outcome for the management of rheumatoid arthritis (RA). Heterogeneity of respective patient-reported outcome measures (PROMs) challenges direct comparisons between their results. This study aimed to standardize reporting of such PROMs measuring functioning in RA to facilitate comparability.

Methods. Common-item nonequivalent group design with the Health Assessment Questionnaire (HAQ) as a common scale across data sets from various countries (including the UK, Turkey, and Germany) to establish a common metric was used. Other PROMs included are the physical function items of the Multidimensional HAQ (MDHAQ), the Disabilities of the Arm, Shoulder, and Hand questionnaire, the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), the World Health Organization Disability Assessment Schedule II (WHODAS II), the Medical Outcomes Study Short Form 36 (SF-36) health survey, and 4 short forms (20, 10, 6, and 4 physical function items) from the Patient-Reported Outcomes Measurement Information System. As the HAQ includes mobility, self-care, and domestic life items, this study focuses on these 3 domains. PROMs were described using standard error of measurement (SEM) and smallest detectable difference (SDD). A Rasch measurement model was used to create the common metric.

Results. The range of the SEM was 0.2 (MDHAQ) to 7.4 (SF-36 health survey physical functioning domain). The SDD revealed a range from 9.7% (WOMAC rating scale) to 33.5% (WHODAS physical functioning domain). PROMs co-calibration revealed fit to the Rasch measurement model. A transformation table was developed to allow exchange between PROM scores.

Conclusion. Scores between the daily activity PROMs commonly used in RA can now be compared. Factors such as SEM and SDD help to determine the choice of a PROM in clinical practice and research.

INTRODUCTION

Studies of the lived experience of individuals with rheumatoid arthritis (RA) show that most facets of life can be affected by the health condition (body structures and functions can be impaired, activities in daily life limited, as well as social, community, and civic life being restricted) (1,2) and thus are important outcomes to measure in evaluating and monitoring the health condition and related interventions. Therefore, a comprehensive understanding

of health, as reflected in a bio-psycho-social perspective, is foundational for measuring outcomes in clinical trials, epidemiologic studies, or the routine monitoring of the patients' progress (3,4). "Outcome" refers here to any indicator (variable) to detect changes in health status or quality of life. Clinicians and researchers use a wide range of outcomes, from inflammatory markers and joint counts through job retention and quality of life (5–8). Many such outcomes use questionnaires to measure patients' perceptions of the condition's impact on their health

and lives. Such patient-reported outcome measures (PROMs) have been used in RA for >35 years (9). In the context of this study, a PROM is defined as any patient- or proxy-completed questionnaire in which a set of items is summated to give a total score, a series of domain scores, or both. "Domain" refers to any meaningful aggregation of categories as defined by the International Classification of Functioning, Disability, and Health (ICF) instrument (10). ICF categories (e.g., d450 Walking) are the unit of the classification and are hierarchically ordered into chapters (e.g., d4 Mobility) and components (e.g., d Activities and Participation). The components and their interactions reflect a bio-psycho-social model of health and disability in RA (11,12).

The use of PROMs in rheumatology is ubiquitous. For example, a recent European Alliance of Associations for Rheumatology (EULAR) PROM program project found that from 2000–2016, 78 different PROMs were used to measure outcomes in osteoarthritis studies (13). Often, several different PROMs can be used to measure the same domain, such as pain, fatigue, mobility, or self-care. This heterogeneity makes it difficult to directly compare the results of PROMs from different studies. Furthermore, data derived from PROMs are often ordinal scaled, limiting their usefulness in monitoring change over time (14). The lack of comparable and interval-scaled information collected from PROMs measuring the same construct restricts using data for secondary clinical purposes, such as quality audits and benchmarking, as well as for research purposes, including meta-analyses. However, international standards for eHealth stress the need for information systems based on international health classifications, including the ICF, to ensure that health information is available in a consistent and comparable manner for effective use in decision-making (15). Therefore, the objective of this study was to standardize reporting of commonly used PROMs in RA to facilitate their comparability.

MATERIALS AND METHODS

Conceptual and score equivalence are foundational to establishing comparability of existing PROMs (16). For conceptual equivalence, we relied on previously linked items from selected PROMs to the ICF (www.icf-research-branch.org). PROMs linked to the same ICF domains are assumed to be comparable from a content perspective and thus could be included in the psychometric analyses to establish score equivalence. The Rasch measurement model was applied, with the total scores of the PROMs equated directly, to establish score equivalence rather than ratings of single items within each PROM (17).

Outcome measures. The 10 most commonly used PROMs in the last 10 years (2006–2016) in RA research were identified based on the preliminary results of the second part of the above mentioned EULAR project focusing on PROMs used in RA. Of those, 6 include items that were linked to the ICF component d Activities and Participation. The remaining 4 were the EuroQol 5-domain instrument (not a summated scale and with mixed domain content) (18), the Hospital Anxiety and Depression scale (19), the 6-dimension Short Form health survey, and the Rheumatoid Arthritis Quality of Life scale (20). The 6 chosen included the Medical Outcomes Study Short Form 36 (SF-36) health survey (21), the Health Assessment Questionnaire (HAQ) (9), the Disabilities of Arm, Shoulder, and Hand (DASH) questionnaire (22), the Multidimensional HAQ (MDHAQ) (23), the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (24), and the World Health Organization Disability Assessment Schedule II (WHODAS II) (25). Other generic PROMs allowing comparability across conditions were included, namely, relevant subscales from the Patient-Reported Outcomes Measurement Information System (PROMIS) (26), as it is a recommended PROM for functional status assessment (27).

Because the HAQ is the most commonly used PROM in RA and covers mainly activities related to mobility, self-care, and domestic life, this study focused on these 3 ICF activities and participation domains. Among the selected PROMs, only (sub-) scales that mapped on to the d4 Mobility, d5 Self-care, and d6 Domestic Life domains were chosen. Since items within each included PROM are generally consistent with undertaking tasks associated with activities of daily living, the resulting interval-scaled common metric was referred to as the "daily activities metric."

The HAQ. The HAQ (9) consists of 20 items assessing difficulties in performing activities of daily living on a scale of 0 = "Without any difficulty" to 3 = "Unable to do." These items are grouped into 8 domains. To create a total score, the highest item scores from each domain are added and then divided by 8, with

higher scores indicating more difficulties. In this study, the HAQ was scored without the score adjustment for assistive devices and help because 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.

SF-36 health survey. The SF-36, version 2 (21), comprises 8 health domains, of which only the physical functioning domain (SF-36-PF) was relevant for this study. The SF-36-PF consists of 10 items related to activities of daily living, each rated on a scale from 1 = "Limited a lot" to 3 = "Not limited at all." The total score is created by summing the responses to each item and transforming it to a 0–100 scale, with lower scores indicating worse function.

The DASH questionnaire. The DASH questionnaire (22) contains 30 items related to physical function and symptoms. Only the 23 items related to physical function (henceforth referred to as the DASH23) were included and rated on a scale from 0 = "No difficulty" to 5 = "Extreme difficulty." The mean of the items is transformed into a scale from 0 to 100 for the total score ([(sum of n responses – 1) / n] × 25), with higher scores indicating worse function.

The MDHAQ. The MDHAQ (23) consists of 10 items: the 8 MHAQ items plus walking 3 kilometers and participating in recreational activities. The total score is the sum of the items divided by the total number of items answered (at least 9 of the 10 are required). The value is rounded to the first decimal, with higher scores indicating worse function.

The WOMAC. The WOMAC (24) consists of 3 subscales (pain, stiffness, and physical function). Only the physical function subscale (WOMAC-PF), which includes 17 items, was included. Two forms of the WOMAC-PF are available; one with a numerical rating scale scored 0–10 (WOMAC NRS), the other with a rating scale scored 0–4, whereby 0 always indicates "no difficulty" and the higher score "extreme difficulty" (WOMAC RAT). Because both forms are used in practice, we included both. A total score for each subscale is created by summing up the respective items, with higher scores indicating worse function.

WHODAS II. The WHODAS II (25) is a generic health and disability instrument with 6 domains. Three domains (mobility, self-care, and life activities) equated to the ICF chapters d4 Mobility, d5 Self-care, and d6 Domestic Life and thus were relevant for this study (henceforth referred to as WHODAS physical functioning domain). Items are scored on a scale from 0 = "No difficulty" to 4 = "Extreme difficulty/cannot do." A total score for each domain is created by summing up its items' responses, with higher scores indicating worse function.

PROMIS. The PROMIS (26) is a set of measures of physical, mental, and social health. In this study, we included the 4 physical function short-forms (PF-20, PF-10, PF-6, and PF-4 items) of the PROMIS. Items are rated on a scale from 1 = "Cannot do" to 5 =

"Without any difficulty." A total score for each short form is created by summing up the responses to each item, with lower scores indicating worse function.

In total, we included 11 PROMs including 4 forms of the PROMIS, 2 forms of the HAQ, and 2 forms of the WOMAC. All PROMs were collected using the validated language version in the participating countries.

Data collection. We adopted a 2-fold strategy. First, we considered data sets in which data of the identified PROMs were already collected previously and applied for data collected in the process of developing and validating the ICF core set for RA at Ludwig-Maximilians-University (LMU), Munich, which coordinated the ICF core set development process relying on an international network. More specifically, we used the data from Lithuania, Serbia, Hungary, and The Netherlands and grouped it together under an "other Europe" label. Participants were diagnosed with RA according to the study criteria of the primary studies.

Second, to ensure that all PROMs, or at least 1 version of each PROM, were well populated in English, German, and Turkish, we collected additional data on individuals with RA at Ankara University, the University of Salford, and LMU, Munich (Figure 1). All relevant documents were prepared in a generic form and then adopted to local regulations by the local research teams to ensure that data collection followed the respective regulations in place at the time. Data collection took place between Spring

Rheumatoid arthritis	HAQ	SF-36-PF	DASH23	WHODAS-PF	PROMIS-SFs	MDHAQ	WOMAC-PF
Ankara (TK)	░		▓	░		▓	▓
Salford (UK)	░	░			▓	▓	
Munich (GER)	░					▓	
Other Europe	░	░					

Figure 1. Overview of the data structure showing previously collected data used in this study for secondary analysis (light gray) and data that have been newly collected specifically for this project (dark gray). DASH23 = 23-item Disabilities of the Arm, Shoulder, and Hand questionnaire; GER = Germany; HAQ = Health Assessment Questionnaire; MDHAQ = Multidimensional Health Assessment Questionnaire; PROMIS-SFs = Patient-Reported Outcomes Measurement Information System short forms; SF-36-PF = Medical Outcomes Study Short Form 36 health survey physical functioning domain; TK = Turkey; WHODAS-PF = World Health Organization Disability Assessment Schedule physical functioning domain; WOMAC-PF = Western Ontario and McMaster Universities Osteoarthritis Index physical functioning domain.

2017 and 2018 through the outpatient clinic or established patient networks at each site. Ethics approval was obtained from the appropriate research ethics committees at each site, and each participant gave his/her informed written consent to participate in this study.

Data analysis. Analysis was embedded within a common-item, nonequivalent group design (NEAT) with the HAQ being the common PROM across all data sets. NEAT implies that the same items were administered in different groups, but not all individuals are administered all items. This design allows bringing together different data sets containing different PROMs but with at least 1 item set common across all sets (the HAQ in the present study). Descriptive statistics were used to describe PROM scores for each country; the Kruskal-Wallis test was used to determine any differences in the ordinal PROM scores across countries. In addition, the standard error of measurement (SEM) ($SD \times \sqrt{[1-\alpha]}$) and smallest detectable difference (SDD) ($SEM \times 1.96 \times \sqrt{2}$) were calculated on the raw scale scores to gain information about the level of precision of the scale. The SDD was also presented as a percentage of the full operational range of the PROM (i.e., its total raw score range). Cronbach's α is reported as an indicator of the internal reliability of each scale.

To co-calibrate the scales onto a common reference metric (an interval-scaled metric with ≥3 scales), the partial credit parameterization of the Rasch measurement model was used within the RUMM2030 software (28,29). The analytical test-equating approach adopted in this study is recent, involving just the total scores of the scales to represent items within the daily activities metric (17). This has the advantage of absorbing any local item dependency that exists within each scale. Thus, the scales intended to measure the daily activities domain were calibrated onto the reference metric, and their fit to the Rasch model tested as a set of items (i.e., each PROM represented an item).

Due to the incomplete nature of the data matrix (not all PROMs were collected in each setting), fit to the model was tested by pairwise PROM fit, with the HAQ always being present. Such a pairwise test of fit makes available a robust conditional test of fit (CTF) to see if the data accord with model expectations (17). Ideal fit values are reported at the bottom of the fit table (see Results section).

Unidimensionality was tested with a principal components analysis of the standardized Rasch residuals. A t-test was conducted comparing pairs of ability estimates, either loading positively or negatively on the first component of the residuals. The lower limit of the confidence interval for the percentage of significant t-tests should be <5%.

Scale invariance was tested by examining differential item functioning (DIF). PROMs were considered as invariant or free of DIF if patients with comparable levels of daily activities ability (as defined by the 2 PROMs under consideration in each pairwise comparison) obtained the same score on a given PROM, regardless of group characteristics (e.g., age, sex, and country). Should DIF be observed, a comparison was made between unadjusted and adjusted person estimates, the latter derived by splitting items on the group variable (29). In this study, if a paired t-test between the 2 estimates was significant, a substantive difference was interpreted as an effect size of that difference ≥0.1 (30).

A core of 6 PROMs, referred to as "core scale bank," was identified and co-calibrated to define the reference metric. This core scale bank was designed to prevent replicates of PROMs (i.e., the 4 PROMIS short forms, the 2 WOMAC physical functioning domain forms, and the HAQ and MDHAQ) to avoid problems with a breach of the local independence assumption and so included the WOMAC RAT, the DASH23 questionnaire, the PROMIS PF-20, the SF-36-PF, the WHODAS physical functioning domain (with its 3 domains summated into a single score), and the HAQ (30). The remaining scales were subsequently calibrated onto the metric on an individual basis, calibrating along with the HAQ, anchored to the item parameters of the HAQ from the core set analysis.

RESULTS

Age, sex, and disease duration of the sample in each country are given in Table 1. The contribution to the overall sample made by each country for each PROM is shown in Table 2. The raw data are presented in the way that they are traditionally reported, for example, variations of the HAQ are rescored to 0–3, and the SF-36-PF to 0–100. Table 3 gives some basic descriptive statistics for each PROM, as well as the SEM and SDD. The WOMAC physical functioning domain (in either format) and the PROMIS 20-item

Table 1. Age, sex, and disease duration of the sample in each country*

Country	Mean ± SD age, years	Female, %	Mean ± SD disease duration, years	Median HAQ score	No.
Germany	49.0 ± 13.8	91.4	13.5 ± 12.2	1.0	180
UK	68.3 ± 10.0	74.2	19.8 ± 13.0	1.0	535
Turkey	57.5 ± 11.5	75.8	13.7 ± 10.3	0.8	458
Other Europe	56.9 ± 12.7	80.4	11.3 ± 9.8	1.5	554
Total	–	78.4	–	–	1,727

* HAQ = Health Assessment Questionnaire.

Table 2. Country contributions to scale bank*

Scale and country	Sample	Median score	IQR	Difference, Kruskal-Wallis P
WHODAS-PF				
Turkey	296	12	4–24	–
SF-36-PF				
UK	368	40	15–65	–
Other Europe	514	35	15–55	0.0283
PROMIS-PF4				
Germany	156	15	12–17	–
UK	152	14	9–18	0.3222
PROMIS-PF6				
Germany	156	20	16–24	–
UK	152	19	12–35	0.2178
PROMIS-PF10				
Germany	156	34	28–41	–
UK	152	32	24–41	0.2788
PROMIS-PF20				
Germany	156	72	60–86	–
UK	152	68	50–85	0.1322
DASH23				
Germany	155	34	19–53	–
Turkey	115	33	16–56	0.7942
WOMAC NRS				
Germany	153	49	20–90	–
WOMAC RAT				
UK	141	24	7–35	–
Turkey	155	20	11–36	0.9170
MDHAQ				
UK	151	1	0.375–1.75	–
Turkey	156	0.63	0.25–1.5	0.0636
HAQ				
Germany	176	1	0.5–1.5	–
UK	529	1	0.5–1.75	–
Turkey	457	0.75	0.125–1.5	–
Other Europe	427	1.5	0.875–2.0	0.0001

* DASH23 = 23-item Disabilities of the Arm, Shoulder, and Hand questionnaire; HAQ = Health Assessment Questionnaire; IQR = interquartile range; MDHAQ = Multidimensional HAQ; NRS = numerical rating scale (range 0–10); PROMIS-PF4 = Patient-Reported Outcomes Measurement Information System 4-item physical functioning domain; RAT = rating scale (range 0–4); SF-36-PF = Medical Outcomes Study Short Form 36 health survey physical functioning domain; WHODAS-PF = World Health Organization Disability Assessment Schedule physical functioning domain; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index.

Table 3. Scale precision ordered by percentage of smallest detectable difference (SDD)*

Scale	No. of observations	Median	IQR	Min	Max	Cronbach's α	SEM	Operational range	SDD	% SDD
WOMAC RAT	296	21.5	9–36	0	85	0.97	2.96	85	8.21	9.66
WOMAC NRS	153	49.3	20.4–90.1	0	170	0.98	6.77	170	18.76	11.03
PROMIS-PF20	298	70.0	54–86	20	100	0.97	3.19	80	8.83	11.04
DASH23	270	33.0	18–55	0	92	0.97	4.02	92	11.14	12.11
PROMIS-PF10	305	33.0	26–41	10	50	0.95	2.21	40	6.13	15.32
MDHAQ	307	0.9	0.38–1.63	0	3	0.91	0.19	3	0.54	17.82
PROMIS-PF6	308	20.0	14–25	6	30	0.94	1.59	24	4.42	18.42
SF-36-PF	882	35.0	15–60	0	100	0.92	7.41	100	20.53	20.53
HAQ	1,589	1.0	0.5–1.75	0	3	0.92	0.23	3	0.63	21.08
PROMIS-PF4	308	14.0	10–18	4	20	0.92	1.33	16	3.68	23.02
WHODAS-PF	296	12.0	4–24	0	52	0.94	6.28	52	17.40	33.47

* % SDD = % of operational range that is the SDD; DASH23 = 23-item Disabilities of the Arm, Shoulder, and Hand questionnaire; HAQ = Health Assessment Questionnaire; IQR = interquartile range; MDHAQ = Multidimensional HAQ; NRS = numerical rating scale; PROMIS-PF4 = Patient-Reported Outcomes Measurement Information System 4-item physical functioning domain; RAT = rating scale; SEM = standard error of measurement; SF-36-PF = Medical Outcomes Study Short Form 36 health survey physical functioning domain; WHODAS-PF = World Health Organization Disability Assessment Schedule physical functioning domain; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index.

Table 4. Fit of scales to the Rasch model*

Fit of the HAQ	Reliability	Conditional test of fit *P*	Unidimensionality *t*-test†	DIF present	Substantive DIF	No.
WOMAC RAT	0.80	0.4504	0.69	Age and country	–	290
WOMAC NRS	0.93	0.1408	1.36	–	–	150
DASH23	0.96	0.9910	3.08	Sex	–	266
PROMIS-PF20	0.98	0.9999	5.44 (LCI 3.7)	–	–	294
PROMIS-PF10	0.96	0.9933	4.01	–	–	299
PROMIS-PF6	0.93	0.8095	3.36	–	–	304
PROMIS-PF4	0.91	0.8601	3.10	–	–	290
SF-36-PF	0.88	0.5783	1.96	Country	–	776
WHODAS	0.75	0.9933	0.37	Age and sex	–	295
Core scale bank	0.87	0.1218	–	Country	–	1,665
Ideal values	>0.7	>0.05	<5.0 (LCI <5.0)	Absent	Absent	–

* DASH23 = 23-item Disabilities of the Arm, Shoulder, and Hand questionnaire; DIF = differential item functioning; HAQ = Health Assessment Questionnaire; LCI = lower confidence interval; NRS = numerical rating scale (range 0–10); PROMIS-PF4 = Patient-Reported Outcomes Measurement Information System 4-item physical functioning domain; RAT = rating scale (range 0–4); SF-36-PF = Medical Outcomes Study Short Form 36 health survey physical functioning domain; WHODAS = World Health Organization Disability Assessment Schedule; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index.
† % at >0.05 (LCI).

physical functioning domain are the most efficient PROMs in that only ~11% of the scale would need to be transited to get above the error. In contrast, the HAQ would need to transit over one-fifth of the PROM (21.1%) or the WHODAS physical functioning domain one-third (33.5 %) to get above the error. In other words, a 15% score change in the HAQ cannot be statistically detected but would be veiled by measurement error, whereas such a change in the WOMAC physical functioning domain can be already detected as statistically significant change.

Fit of the PROMs to the Rasch model is shown in Table 4. The 4 PROMIS physical functioning domain sets and the SF-36 physical functioning domain had their scores reversed to be consistent with the other PROMs, so that a high score indicates poor functioning. Each row is a pairwise fit of the HAQ plus one other scale until the final row brings together a number of scales (core set), avoiding putting scales together that are close replicates of one another. All pairs of scales showed fit to the Rasch model, represented by a non-significant CTF, and all pairs were unidimensional. Some DIF was observed and tested to see if this gave rise to significantly different person estimates. Substantive DIF was absent at the pairwise level; for example, the WOMAC rating scale showed a paired *t*-test significance of 0.83. In the 6-PROM core scale bank, the country-based DIF for the WOMAC rating scale was still present. Nevertheless, the effect size of the differences (between the unadjusted and adjusted analyses) was just 0.07, and thus considered to be negligible (31).

Given that all the PROMs tested fit, the assumptions of the Rasch model, a transformation table was created. Supplementary Table 1, available on the *Arthritis Care & Research* website at http://onlinelibrary.wiley.com/doi/10.1002/acr.24503/abstract, shows the exchange rates between the 11 PROMs tested (i.e., including the 4 forms of the PROMIS; the HAQ and MDHAQ;

and 2 forms of the WOMAC-PF) using the interval-scaled daily activities metric as the link. A high score on this reference metric represents low ability to perform tasks and, conversely, a low score represents high ability. The HAQ and MDHAQ were scored in their usual way of 0–3, and the 4 PROMIS short forms and the SF-36 physical functioning domain scores were reversed so that a high score represented few, if any, limitations in daily activities. For example, a HAQ score of 0.75 was associated with a reference metric score of 43.44, as were a WOMAC rating scale score of 17,
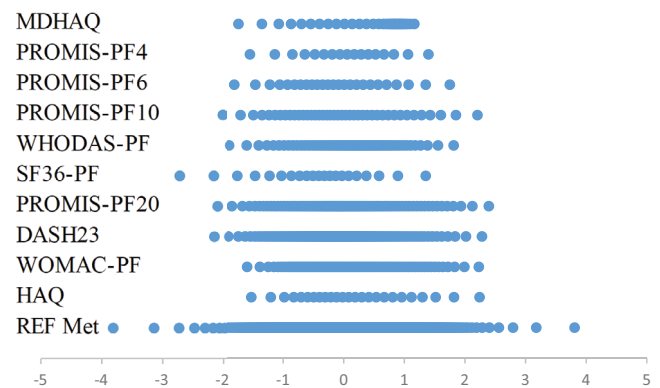

Figure 2. Operational widths of scales on the interval-scaled daily activities metric. DASH23 = 23-item Disabilities of the Arm, Shoulder, and Hand questionnaire; HAQ = Health Assessment Questionnaire; MDHAQ = Multidimensional Health Assessment Questionnaire; PROMIS-PF4 = Patient-Reported Outcomes Measurement Information System 4-item physical functioning domain; REF Met = reference metric; SF-36-PF = Medical Outcomes Study Short Form 36 health survey physical functioning domain; WHODAS-PF = World Health Organization Disability Assessment Schedule physical functioning domain; WOMAC-PF = Western Ontario and McMaster Universities Osteoarthritis Index physical functioning domain.

a DASH23 score of 28, and an SF-36 physical functioning domain score of 55. If there was no direct match, the nearest score was taken (e.g., a PROMIS physical functioning domain score of 20 of 77, and a WHODAS physical functioning domain score of 13). Even where there is no direct match, the link will be accurate within less than one-tenth of a logit. To facilitate access to the reference metric, Supplementary Table 1 (available at http://onlinelibrary. wiley.com/doi/10.1002/acr.24503/abstract) is presented as an Excel (Microsoft) supplementary file. Thus, readers can choose to select just those PROMs relevant to their current analysis to obtain the interval-scaled daily activities metric or compare PROM scores, or both.

Figure 2 shows the operational ranges of the scales in logits along with the interval-scaled daily activities metric. Most scales measure a similar range, i.e., within ±2 logits, with only slight variations. These variations manifest also in the transformation Table (see Supplementary Table 1, available at http://online library.wiley.com/doi/10.1002/acr.24503/abstract) where, for example, the SF-36 physical functioning domain has the lowest reference metric of all the scales, with 14.20 for its score of 100, but only achieves a metric level of 67.45 for its score of 0. Thus, its orientation is slightly to the more able end of the daily activities metric.

DISCUSSION

Many of the most widely used PROMs in RA involve the measurement of activities of daily living, sometimes referred to as physical function, and are consistent with ICF chapters d4 Mobility, d5 Self-care, and d6 Domestic Life. In this study, 11 PROMs were shown to map onto a daily activities metric, and each pair of PROMs, with the HAQ as a common PROM comparator, showed fit to the Rasch model and unidimensionality. A core set of 6 PROMs also showed such fit. Given that the PROMs all measure a similar range on the daily activities metric, other factors such as the SEM and SDD can be used to differentiate between PROMs when selecting which to use in clinical practice or research. For example, the selected items of the DASH23 for upper limb therapy and research, the WOMAC rating scale version for lower limbs, and the PROMIS 20-item physical functioning domain for general use would seem to be the better choices among these PROMs. Of particular concern is the SDD of the HAQ, indicating that it is less than optimal for detecting a difference compared to other PROMs.

The approach to use just the total scores of the PROMs as items to fit the Rasch model is relatively new (17). Under the Rasch model, sufficiency is explicitly on the total score of the person for the person parameter, and the total score for the item parameter (32). Here, the item is a PROM, and thus the total score for the PROM (summed over all persons) estimates the scale parameter. Increasingly, studies are published that examine the potential of standardized reporting by linking

commonly used questionnaires (33,34). The present study differs from these studies, as the calibration model used here delivers estimates that are independent of the distribution upon which the calibration is based. Such a calibration model requires parameter separation between persons and items (35), which is consistent with applying the Rasch model, as in the current study. Under these circumstances, and given the same frame of reference (e.g., health condition group), clinicians and researchers can have confidence that the transformations (by using, for example, a transformation table) apply to their own sample, involving the same frame of reference. Nevertheless, given the availability of different studies linking commonly used questionnaires to enable comparability using item response theory (IRT) and Rasch models, it remains to be investigated to compare the performance of these different approaches.

The limitations of the study arise from a number of technical issues related to the application and interpretation of the results. For example, current software constraints limit the operational range of an item (in the case of RUMM2030) to 100 categories. Thus, the WOMAC numerical rating scale, with a range of 170, had to be divided by 1.7 and rounded for fit to the model and then expanded again for comparability purposes. The use of the transformation Table (see Supplementary Table 1, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24503/abstract) itself is also constrained to where there are complete data, although recent work has shown that if necessary, imputation of missing data (missing completely at random or at random) will not affect the interpretation of fit to the Rasch model (36). Missing data at the scale level are treated in the same way as in item-based analysis; that is, estimates are based on the information available (i.e., the scale is treated as missing for that case), but missing data are always an indicator of the validity of the scale in a given population irrespective of the analytical strategy chosen. The sample size, while adequate for the Rasch model application, nevertheless is modest compared with other equating studies using different IRT approaches (33,34), but the latter require much larger sample sizes for their chosen models.

The strengths of the study come from the content comparability checks based on the ICF and the confirmation of unidimensionality of the item sets through the Rasch model. The model itself has sufficiency of the person score, such that the only information required is the total score for the person (32). Thus, clinicians and researchers can simply add up the responses to a set of items and have access to the daily activities metric through the transformation table (see Supplementary Table 1, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24503/abstract). The link to the ICF is also consistent with the latest requirements for eHealth informatics, such that data are recorded based on international standards, with the ICF being one of these (15). As such, the approach supports standardized reporting, as there is no need to create new PROMs unless there is a sound reason for doing so (e.g., poor psychometrics in the target population).

The scale banking also facilitates comparability of data and results of clinical trials (e.g., through meta-analysis) and patient registries.

In conclusion, many scales used to assess the impact of RA involve PROMs that ascertain the level of difficulty across a range of everyday activities as described in chapter d4 Mobility, d5 Self-care, and d6 Domestic Life in the ICF. Data from a mix of the most commonly used PROMs in RA have shown that they consistently map onto these chapters. Fit of their data to the Rasch model has shown that in a pairwise fashion, and with a core set of 6 PROMs, the data satisfied the Rasch model expectations, making their total scores comparable via an interval-scaled daily activities metric. Descriptive analysis of the scales suggested that, given similar operational ranges on the metric, some PROMs displayed much lower SDDs in relation to their operational range, which will have implications for sample size requirements and detection of change.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Prodinger had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study conception and design. Prodinger, Tennant.
Acquisition of data. Coenen, Hammond, Küçükdeveci.
Analysis and interpretation of data. Prodinger, Tennant.

REFERENCES

1. Stack RJ, van Tuyl LH, Sloots M, van de Stadt LA, Hoogland W, Maat B, et al. Symptom complexes in patients with seropositive arthralgia and in patients newly diagnosed with rheumatoid arthritis: a qualitative exploration of symptom development. Rheumatology (Oxford) 2014; 53:1646–53.

2. Sverker A, Östlund G, Thyberg M, Thyberg I, Valtersson E, Björk M. Dilemmas of participation in everyday life in early rheumatoid arthritis: a qualitative interview study (the Swedish TIRA Project). Disabil Rehabil 2014;37:1251–9.

3. Nicassio PM, Kay MA, Custodio MK, Irwin MR, Olmstead R, Weisman MH. An evaluation of a biopsychosocial framework for health-related quality of life and disability in rheumatoid arthritis. J Psychosom Res 2011;71:79–85.

4. Coenen M, Cieza A, Stamm TA, Amann E, Kollerits B, Stucki G. Validation of the International Classification of Functioning, Disability and Health (ICF) core set for rheumatoid arthritis from the patient perspective using focus groups. Arthritis Res Ther 2006;8:R84.

5. Gilworth G, Chamberlain MA, Harvey A, Woodhouse A, Smith J, Smyth MG, et al. Development of a work instability scale for rheumatoid arthritis. Arthritis Rheum 2003;49:349–54.

6. Carr A, Hewlett S, Hughes R, Mitchell H, Ryan S, Carr M, et al. Rheumatology outcomes: the patient's perspective. J Rheumatol 2003;30: 880–3.

7. Kalyoncu U, Dougados M, Daurès J, Gossec L. Reporting of patient-reported outcomes in recent trials in rheumatoid arthritis: a systematic literature review. Ann Rheum Dis 2009;68:183–90.

8. Heller J, Shadick NA. Outcomes in rheumatoid arthritis: incorporating the patient perspective. Curr Opin Rheumatol 2007;19:101–5.

9. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.

10. WHO. International Classification of Functioning, Disability and Health. Geneva: World Health Organization (WHO); 2001.

11. McCollum L, Pincus T. A biopsychosocial model to complement a biomedical model: patient questionnaire data and socioeconomic status usually are more significant than laboratory tests and imaging studies in prognosis of rheumatoid arthritis. Rheum Dis Clin North Am 2009;35:699–712.

12. Stucki G, Cieza A, Geyh S, Battistella L, Lloyd J, Symmons D, et al. ICF core sets for rheumatoid arthritis. J Rehabil Med 2004 Suppl 44: 87–93.

13. Lundgren-Nilsson Å, Dencker A, Palstam A, Person G, Horton MC, Escorpizo R, et al. Patient-reported outcome measures in osteoarthritis: a systematic search and review of their use and psychometric properties. RMD Open 2018;4:e000715.

14. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med 2012;44:97–8.

15. ISO. Health informatics – capacity-based eHealth architecture roadmap – part 2: architectural components and maturity model. PD ISO/TR 14639-2:2014. UK: International Standard Organisation; 2014.

16. 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 Policy 2016;21:223–8.

17. 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: Institute for Learning Sciences and Teacher Education; 2013. p. 164–96.

18. Hurst N, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H, for the Economic and Health Outcomes Research Group. Validity of EuroQoL—a generic health status instrument—in patients with rheumatoid arthritis. Br J Rheumatol 1994;33:655–62.

19. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67:361–70.

20. De Jong Z, van Der Heijde D, McKenna SP, Whalley D. The reliability and construct validity of the RAQoL: a rheumatoid arthritis specific quality of life instrument. Br J Rheumatol 1997;36:878–83.

21. Ware JE, Kosinski M, Dewey JE, Gandek B. How to score version 2 of the SF-36® Health Survey. Lincoln (RI): Quality Metric; 2000.

22. Hudak PL, Amadio PC, Bombardier C, Beaton D, Cole D, Davis A, et al. Development of an upper extremity outcome measure: the DASH (Disabilities of the Arm, Shoulder, and Hand). Am J Ind Med 1996;29:602–8.

23. Pincus T, Swearingen C, Wolfe F. Toward a Multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42: 2220–30.

24. Bellamy N. The WOMAC knee and hip osteoarthritis indices: development, validation, globalization and influence on the development of the AUSCAN hand OA indices. Clin Exp Rheumatol 2005;23:S148.

25. World Health Organization. WHO disability assessment schedule 2.0 (WHODAS2.0). Geneva: World Health Organization; 2013. URL: http://www.who.int/classifications/icf/whodasii/en/index.html.

26. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol 2010;63: 1179–94.

27. Barber CE, Zell J, Yazdany J, Davis AM, Cappelli L, Ehrlich-Jones L, et al. 2019 American College of Rheumatology recommended patient-reported functional status assessment measures in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2019;71:1531–9.

28. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149–74.

29. Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. Health Qual Life Outcomes 2017;15:181.

30. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas 2008;9:200–15.

31. Rouquette A, Hardouin JB, Vanhaesebrouck A, Sébille V, Coste J. Differential item functioning (DIF) in composite health measurement scale: recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. PLoS One 2019;14: e0215073.

32. Andersen EB. Sufficient statistics and latent trait models. Psychometrika 1977;42:69–81.

33. 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;24:2305–18.

34. Oude Vsohaar MA, Vonkeman HE, Courvoisier D, Finckh A, Gossec L, Leung YY, et al. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Life Res 2019;28:187–97.

35. Andrich D. Rasch models for measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences. Newbury Park (CA): Sage Publications; 1988.

36. Fellinghauer C, Prodinger B, Tennant A. The impact of missing values and single imputation upon Rasch analysis outcomes: a simulation study. J Appl Meas 2018;19:1–25.