Rheumatology, 2025, **00**, 1–11
https://doi.org/10.1093/rheumatology/keaf189
Advance access publication 22 April 2025
Original Article

British Society for Rheumatology RHEUMATOLOGY OXFORD

Clinical science

Patient Reported Outcome Measures for Rheumatoid Arthritis Disease Activity: Rasch measurement theory to identify items and domains

Tim Pickles [1,*], Mike Horton [2], Karl Bang Christensen [3], Rhiannon Phillips [4], David Gillespie [1], Neil Mo [5], Janice Davies [6], Susan Campbell [6], Ernest Choy [7]

[1]Centre for Trials Research, Cardiff University, Cardiff, UK
[2]Psychometric Laboratory for Health Sciences, University of Leeds, Leeds, UK
[3]Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark
[4]Cardiff School of Sport and Health Sciences, Cardiff Metropolitan University, Cardiff, UK
[5]Swansea Bay University Health Board, Port Talbot, UK
[6]Patient and Public Involvement Stakeholders, UK
[7]Division of Infection and Immunity, Cardiff University, Cardiff, UK

*Correspondence to: Tim Pickles, Centre for Trials Research, Cardiff University, 5th Floor, Neuadd Meirionnydd, Heath Park, Cardiff CF14 4YS, UK.
E-mail: PicklesTE@cardiff.ac.uk

Abstract

Objectives: Disease activity (DA) monitoring is a standard of care in RA. There is demand for achieving this through patient-reported outcome measures (PROMs). The aim of this study was to determine which items could be used to measure the construct of RA DA, by analysing legacy PROMs, using Rasch measurement theory (RMT) analyses.

Methods: Questionnaires including 10 legacy PROMs were sent to people with RA to create original and validation datasets. Items were grouped according to OMERACT domains and analysed using principal component analysis. Based on separate domain RMT analyses of the original dataset, domain-level testlets were assessed to determine which items measure the construct of RA DA. The result was then replicated in confirmatory factor analyses bifactor models and RMT analyses of the validation dataset. Psychometric properties of legacy PROMs were also assessed in the original dataset.

Results: The total sample size was 691 (original: 398, validation: 293). The *Patient global* domain was split into *General health* and *Disease activity* domains under RMT. *General health* and *Fatigue* domain items measure a separate construct to the construct of RA DA. A set of 12 *Pain*, *Disease activity*, *Tenderness and swelling*, *Physical functioning* and *Stiffness* domain items can be used to measure the construct of RA DA. No legacy PROMs fully fit the Rasch measurement model.

Conclusion: *General health* and *Disease activity* domain items are not interchangeable. Twelve items form an item pool that can be used to measure the construct of RA DA. Legacy PROMs should not be recommended for use.

Keywords: rheumatoid arthritis disease activity, patient-reported outcome measures, measurement properties.

Rheumatology key messages

- General health and disease activity (DA) domain items are not interchangeable.
- RA DA requires use of the *Tenderness and swelling*, *Pain*, *Disease activity*, *Stiffness* and *Physical functioning* domain items.
- No legacy PROMs fully fit the Rasch measurement model.

Introduction

Patient-reported outcome measures (PROMs) are critical to research and clinical care, as recognized by the U.S. Food and Drug Administration (FDA), who mandated PROMs to be captured in all randomized controlled trials. Additionally, the FDA have published guidelines on how to develop and validate PROMs [1, 2]. Disease activity (DA) monitoring is a standard of care in RA, and there is demand for achieving this through PROMs. Although there are many RA DA PROMs [1], these are currently used as secondary outcomes in clinical trials of rheumatic diseases, but rarely in clinical care. All of the PROMs were developed using classical test theory

Revised: 19 February 2025. **Accepted:** 31 March 2025

methods and many have various limitations. The FDA [2, 3] and COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines [4–6] both recognize item response theory (IRT) and Rasch measurement theory (RMT) as suitable methods for assessing the measurement properties of instruments. Validation using these methods requires that PROMs meet stringent measurement criteria, which include unidimensionality, internal consistency, targeting, lack of local dependence, and differential item functioning (DIF). Thus, IRT and RMT provide a statistical framework within which all these measurement criteria can be formulated as testable hypotheses. Specifically, RMT [7–9] allows for these attributes to be formally assessed, as it provides a template for determining PROM score validity.

A systematic review [10] of 10 legacy RA DA PROMS showed that none can be recommended for use according to COSMIN guidelines [4–6]. This justifies the collection of further data to start the process of determining the domains, and items within those domains, that can be used to measure the construct of RA DA.

The overall aim of this study was to use RMT analyses to determine which items can form an item pool for measuring the construct of RA DA. A secondary aim was to examine the measurement properties of legacy RA DA PROMs and other relevant PROMs.

Methods

This research is reported in line with the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) framework (Supplementary Data S1, available at *Rheumatology* online) [11].

Study design

This was a cross-sectional study that took place in 2020 and 2021. In Cardiff and Vale University and Swansea Bay University Health Boards (UHBs), potential participants were identified by NHS staff by searching the electronic health records of the Rheumatology Department for those patients at least 18 years old with RA. In Aneurin Bevan UHB, potential participants were identified by NHS staff as those at least 18 years old with an entry on the British Society for Rheumatology Biologics Registry for Rheumatoid Arthritis (BSRBR-RA) database. Paper questionnaires were sent out as part of study packs to these people living with RA (plwRA). In Cwm Taf Morgannwg UHB, NHS staff identified potential participants as those patients with RA in the clinic who were at least 18 years old, and they handed them the study pack. Inclusion criteria were: being at least 18 years of age, having a diagnosis of RA, and providing signed informed consent. Patients were excluded if they were unable to complete the questionnaire in English. The study was approved by the North West—Preston Research Ethics Committee (20/NW/0039).

Sample size

To provide item calibrations within ±0.5 logits within a RMT analysis, the advised sample size is 250 [12]. Given this, it was decided that a sample size of $n \geq 250$ was required, for both an original dataset and a validation dataset.

Questionnaire creation

A questionnaire was created based on the items from 10 legacy PROMs identified and reviewed in a systematic review (see Supplementary Data S2, available at *Rheumatology* online) [10]:

- Rheumatoid Arthritis Disease Activity Index-5 (RADAI5) [13–15];
- Rheumatoid Arthritis Disease Activity Index (RADAI) [16, 17];
- RADAI-SF [17, 18];
- Patient-based Disease Activity Score 2 (PDAS2) [19, 20];
- Patient-reported Outcome CLinical ARthritis Activity (PRO-CLARA) [21];
- Global Arthritis Score (GAS) [22];
- Patient Activity Score (PAS) [23];
- Patient Activity Score-II (PAS-II) [23];
- Routine Assessment of Patient Index Data 3 (RAPID3) [24];
- Routine Assessment of Patient Index Data 4 (RAPID4) [25].

Also included were the items from two PROMs measuring level of flare:

- Rheumatoid Arthritis Flare Questionnaire (RA-FQ) [26, 27];
- FLARE-RA (which includes FLARE-RA Old, FLARE-RA Arthritis and FLARE-RA General Symptoms) [28–31].

The items of The Rapid Assessment of Disease Activity in Rheumatology (RADAR) [32, 33], the PROM-score [34] and the foot-specific RADAI-F5 [35], were included, as were fatigue items included on the PAS and PAS-II assessments, the HAQ (PDAS2, PAS) and the multidimensional HAQ (MDHAQ) (used in RAPID3, RAPID4). The HAQ also has an additional pain item. RA-FQ has additional items about having a flare and how long it has been going on.

A draft questionnaire containing these items was discussed with two groups of plwRA: in a meeting with J.D. and S.C., and with a focus group convened by the National Rheumatoid Arthritis Society (NRAS). From these discussions, items on discomfort when walking, standing and exercising, plus fear of falling when walking were added. These four items used the Copenhagen Hip and Groin Outcome Score (HAGOS) [36] as a template. A focus group attendee also provided a pain scale, which was included. Thus, the total item pool contained 268 items (see Supplementary Data S2, available at *Rheumatology* online, which states item codes).

Demographic items relating to current age, age at diagnosis, gender and sex assigned at birth, shielding during the COVID-19 pandemic, whether the participant completed the questionnaire themselves, ethnicity, education level, earlier and accompanying diseases, current or previous DMARD treatment were also included.

Item grouping

All items in the questionnaire, minus the two homunculi (`G01`, `A02`) and the aids and devices and help from another person items from HAQ (`H10`, `H11`, `H23`, `H24`), were grouped according to OMERACT domains for RA [37, 38].

Table 1. Items grouped by OMERACT domain

OMERACT domain	Number of items
Tenderness and swelling	3
Patient global	15
Pain	11
Pain (area-specific)	53
Fatigue	5
Physical functioning	5
Physical functioning (specific)	40
Stiffness	5
Swelling	1
Discomfort/fear	4
Mood	3

The 145 items were initially grouped by T.P. (researcher) and then checked by E.C. (rheumatologist) to ensure correct grouping. Where necessary, additional domains were created (Table 1).

Analyses

Principal component analysis—original dataset only

Principal component analyses (PCAs) [39] were undertaken on the 145 items described listed in Table 1. Two PCAs were undertaken, one using a polychoric correlation matrix and another using Pearson's correlation coefficients. Within the PCA, the principal-component factor method was used, and only factors with a minimum eigenvalue of 1 were retained. Oblique promax rotation was then applied. The purpose was to see whether items within the identified domains loaded together onto factors that reflected those domains. If this was the case, the domain and the items loading to that domain were carried forward to further RMT analyses.

Rasch measurement theory—original and validation datasets

The Rasch Measurement Model (RMM) is a statistical model [7–9, 40] in which the sum score of the item responses contains all information about the underlying latent trait, here the construct of RA DA, in a statistical concept known as sufficiency. The satisfaction of RMM assumptions, therefore, provides a prescription for what is necessary for a PROM to deliver fundamental measurement [41].

Items were assessed by RMT analyses, which provides results on targeting and item locations, overall and individual item fit to the RMM, internal consistency, local dependency, unidimensionality, and item threshold ordering. DIF was investigated by age group (18–54, 55–74, 75+ years), age at diagnosis (2–36, 37–56, 57+ years), sex (male, female), earlier and accompanying diseases (yes, no), previous DMARD treatment (yes, no), and highest educational qualification (qualifications below university graduate, university graduate qualification as minimum). Grouping for age group and age at diagnosis were determined by the interquartile ranges for these variables.

RMT analyses in the original dataset were undertaken on items grouped by domain, with the purpose of identifying potential items within each domain that were candidate items for an item pool.

In the validation dataset, RMT analyses were undertaken on the potential items for each domain. Where discrepancies were found, these were reported. If suitable, items within domains were grouped together to form domain-level testlets, which operate as single items that represent a domain. These domain-level testlets were assessed together by RMT analyses to determine whether they could measure the construct of RA DA. If any evidence was found that this was not the case, iterative changes were made to achieve better fit to the RMM.

Structural validity—original and validation datasets

A confirmatory factor analysis (CFA) model is a statistical model used to test whether measures of a construct are consistent with a hypothesized measurement model based on theory and/or previous analytic research [42, 43]. CFA using Mplus [44] was used to calculate a X^2-test, root mean square error of approximation (RMSEA) along with an accompanying 90% CI, comparative fit index (CFI), Tucker–Lewis index (TLI), standardized root mean square residual (SRMR), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

CFA was applied to the validation dataset to examine whether the solution determined by RMT analyses could replicated in CFA using bifactor models [45].

Legacy PROMs—original dataset only

To assess construct validity, Mann–Whitney U tests [46] were performed to see whether there was a difference between those identifying as having a flare and not having a flare, with a Hodges–Lehmann median difference and 95% CI being calculated [47]. Spearman's ρ correlation coefficients [48] were calculated between legacy PROM scores, with the hypothesis that all ρ were ≥ 0.5. To assess internal consistency, Cronbach's α [49] values were calculated. In line with COSMIN guidelines [4–6], internal consistency was indicated by α being >0.7. Legacy PROMs in the original dataset were assessed using CFA. In line with COSMIN guidelines [4–6], structural validity was indicated by RMSEA being <0.06, TLI being >0.95, CFI being >0.95 and SRMR being <0.08. RMT analyses were applied to the legacy PROMs in the original dataset to assess the structural validity, internal consistency, and measurement invariance measurement properties.

Results

Descriptives

The total sample size was $n = 691$, with $n = 398$ in the original dataset and $n = 293$ in the validation dataset. Study packs were sent out in batches in September 2020 and June, October and November 2021. The mean current age was 63.8 (s.d. 12.82) years, the mean age at diagnosis was 46.4 (s.d. 15.69) years, and the mean disease duration was 17.3 (s.d. 13.65) years. 67.4% (466/691) were female and all were the same as assigned at birth (Table 2). 15.5% (107/691) completed all demographic questions and legacy PROM items of the questionnaire.

Principal component analysis—original dataset

From the results of both PCAs, a set of 30 items loaded together with other items in the domains they were grouped in, a priori. These were taken forward for RMT analyses. These items were in the *Tenderness and swelling*, *Patient global*, *Pain*, *Fatigue*, *Physical functioning* and *Stiffness* domains (Fig. 1).

Table 2. Descriptives of the sample of people living with RA who responded to the questionnaire

		Dataset						Total		
		Original			Validation					
		n	%/Mean	S.D.	*n*	%/Mean	S.D.	*n*	%/Mean	S.D.
Current age		397	63.6	13.25	292	64.0	12.23	689	63.8	12.82
Age at diagnosis		383	45.9	15.68	283	47.1	15.71	666	46.4	15.69
Disease duration		382	17.6	13.82	283	16.9	13.42	665	17.3	13.65
Gender	Male	122	30.7		103	35.2		225	32.6	
	Female	276	69.3		190	64.8		466	67.4	
	Prefer to self-describe	0	0.0		0	0.0		0	0.0	
	Rather not say	0	0.0		0	0.0		0	0.0	
Same gender as assigned at birth?	Yes	398	100.0		292	100.0		690	100.0	
	No	0	0.0		0	0.0		0	0.0	
	Rather not say	0	0.0		0	0.0		0	0.0	
Have you received a shielding letter from the Welsh Government or NHS?	Yes	325	81.7		216	74.0		541	78.4	
	No	70	17.6		75	25.7		145	21.0	
	Don't know	3	0.8		1	0.3		4	0.6	
	Rather not say	0	0.0		0	0.0		0	0.0	
Completed questionnaire on behalf?	Yes	19	4.8		25	8.6		44	6.4	
	No	376	95.2		266	91.4		642	93.6	
	Rather not say	0	0.0		0	0.0		0	0.0	
Best description of ethnic group or background	White (English/Welsh/Scottish/Northern Irish/British)	361	91.2		277	95.2		638	92.9	
	White—other	12	3.0		11	3.8		23	3.3	
	Black/African/Caribbean/Black British	4	1.0		1	0.3		5	0.7	
	Asian/Asian British	10	2.5		0	0.0		10	1.5	
	Mixed/multiple ethnic groups	7	1.8		1	0.3		8	1.2	
	Other	2	0.5		1	0.3		3	0.4	
	Rather not say	0	0.0		0	0.0		0	0.0	
Highest educational qualification?	Usual high school qualifications in your country at age 16 (e.g. GCSE, O-Level)	92	23.1		90	30.9		181	26.4	
	Usual high school qualifications in your country at age 18 (e.g. AS Level, A-Level)	25	6.3		18	6.2		43	6.3	
	A college or university diploma or degree	132	33.5		96	33.0		228	33.3	
	A higher degree or professional qualification (e.g. Doctorate or Masters level degree)	54	13.7		32	11.0		86	12.6	
	None of these qualifications	58	14.7		31	10.7		89	13.0	
	Other	30	7.6		21	7.2		51	7.4	
	Rather not say	4	1.0		3	1.0		7	1.0	
MTX—previous treatment	Yes	154	39.0		118	41.0		272	39.8	
MTX—current treatment	Yes	211	53.4		135	46.9		346	50.7	
SSZ—previous treatment	Yes	156	39.5		105	36.5		261	38.2	
SSZ—current treatment	Yes	85	21.5		63	21.9		148	21.7	
HCQ—previous treatment	Yes	83	21.0		49	17.0		132	19.3	
HCQ—current treatment	Yes	102	25.8		59	20.5		161	23.6	
LEF—previous treatment	Yes	44	11.1		31	10.8		75	11.0	
LEF—current treatment	Yes	19	4.8		10	3.5		29	4.2	
Prednisolone—previous treatment	Yes	121	30.6		94	32.6		215	31.5	

(continued)

Table 2. (continued)

		Dataset				Total	
		Original		Validation			
		n	%/Mean|S.D.	*n*	%/Mean|S.D.	*n*	%/Mean|S.D.
Prednisolone—current treatment	Yes	85	21.5	39	13.5	124	18.2
Enbrel/benepali (etanercept)—previous treatment	Yes	45	11.4	51	17.7	96	14.1
Enbrel/benepali (etanercept)—current treatment	Yes	38	9.6	33	11.5	71	10.4
Humira/amgevita (adalimumab)—previous treatment	Yes	36	9.1	27	9.4	63	9.2
Humira/amgevita (adalimumab)—current treatment	Yes	25	6.3	24	8.3	49	7.2
Cimzia (certolizumab)—previous treatment	Yes	16	4.1	5	1.7	21	3.1
Cimzia (certolizumab)—current treatment	Yes	11	2.8	1	0.3	12	1.8
Remicade/inflectra (infliximab)—previous treatment	Yes	13	3.3	16	5.6	29	4.2
Remicade/inflectra (infliximab)—current treatment	Yes	4	1.0	4	1.4	8	1.2
Simponi (golimumab)—previous treatment	Yes	0	0.0	2	0.7	2	0.3
Simponi (golimumab)—current treatment	Yes	0	0.0	0	0.0	0	0.0
Orencia (abatacept)—previous treatment	Yes	14	3.5	9	3.1	23	3.4
Orencia (abatacept)—current treatment	Yes	9	2.3	7	2.4	16	2.3
Mabthera (rituximab)—previous treatment	Yes	24	6.1	19	6.6	43	6.3
Mabthera (rituximab)—current treatment	Yes	26	6.6	16	5.6	42	6.1
Roactemra (tocilizumab)—previous treatment	Yes	13	3.3	12	4.2	25	3.7
Roactemra (tocilizumab) – current treatment	Yes	17	4.3	9	3.1	26	3.8
Kevzara (sarilumab)—previous treatment	Yes	1	0.3	0	0.0	1	0.1
Kevzara (sarilumab)—current treatment	Yes	2	0.5	0	0.0	2	0.3
Xeljanz (tofacitinib)—previous treatment	Yes	2	0.5	3	1.0	5	0.7
Xeljanz (tofacitinib)—current treatment	Yes	2	0.5	1	0.3	3	0.4
Olumiant (baricitinib)—previous treatment	Yes	11	2.8	12	4.2	23	3.4
Olumiant (baricitinib)—current treatment	Yes	15	3.8	22	7.6	37	5.4
FM	Yes	25	6.5	24	8.6	49	7.4
OA	Yes	127	33.2	83	29.6	210	31.7
Cancer	Yes	48	12.5	32	11.4	80	12.1
Heart disease	Yes	49	12.8	32	11.4	81	12.2
Chronic bronchitis	Yes	20	5.2	11	3.9	31	4.7
Depression	Yes	69	18.0	50	17.9	119	17.9
Diabetes	Yes	44	11.5	30	10.7	74	11.2
Stroke	Yes	15	3.9	13	4.6	28	4.2
Other medical condition	Yes	173	45.2	124	44.3	297	44.8
Site	Cardiff and Vale UHB	308	77.4	1	0.3	309	44.7
	Swansea Bay UHB	69	17.3	275	93.9	344	49.8
	Aneurin Bevan UHB	20	5.0	10	3.4	30	4.3
	Cwm Taf Morgannwg UHB	1	0.3	7	2.4	8	1.2

(continued)

Table 2. (continued)

		Dataset				Total	
		Original		Validation			
		n	%/Mean|S.D.	*n*	%/Mean|S.D.	*n*	%/Mean|S.D.
In addition to Sex (via Gender and Same as assigned at birth?), the following variables are used for the assessment of differential item functioning under RMT							
Age group, years	18–54	85	21.4	64	21.9	149	21.6
	55–74	240	60.5	170	58.2	410	59.5
	75+	72	18.1	58	19.9	130	18.9
Age at diagnosis group, years	2–36	109	28.6	72	25.4	181	27.3
	37–56	166	43.6	130	45.9	296	44.6
	57+	106	27.8	81	28.6	187	28.2
Earlier and accompanying diseases	Yes	292	76.2	214	76.4	506	76.3
	No	91	23.8	66	23.6	157	23.7
Previous DMARD treatment	Yes	292	73.9	215	74.7	507	74.2
	No	103	26.1	73	25.3	176	25.8
Highest educational qualification	Qualifications below university graduate	204	52.3	160	55.6	364	53.7
	University graduate qualification as minimum	186	47.7	128	44.4	314	46.3

UHB: University Health Board; RMT: Rasch measurement theory.

Figure 1. Principal component analyses summary

Table 3. Details from the *Patient global* domain RMT analysis: residual principal component loading and residual correlations

Item	Domain	Residual loading on first principal component	Residual correlations								
			T01	D01	Q03	PS1	A01	P01	R05	PS2	T04	C01
T01	*Disease activity*	0.768										
D01		0.749	0.548[a]									
Q03		0.694	0.573[a]	0.420[a]								
PS1		0.407	0.043	0.205[a]	0.022							
A01		0.264	−0.054	0.088	−0.100	0.426[a]						
P01	*General health*	−0.278	−0.323	−0.365	−0.284	−0.161	−0.157					
R05		−0.328	−0.307	−0.372	−0.393	−0.215	−0.196	−0.020				
PS2		−0.605	−0.430	−0.296	−0.375	−0.120	−0.171	−0.040	−0.145			
T04		−0.693	−0.398	−0.381	−0.347	−0.356	−0.213	−0.129	−0.033	0.467[a]		
C01		−0.717	−0.441	−0.449	−0.364	−0.326	−0.211	−0.048	−0.025	0.405[a]	0.565[a]	

[a] Indicates correlations above the threshold for local dependence of (mean residual correlation + 0.2) = (−0.1 + 0.2) = 0.1.

Rasch measuement theory—original dataset

Tenderness and swelling

The three items (`D02`, `T02`, `Q04`) in the *Tenderness and swelling* domain provided good fit to the RMM and were retained.

Patient global

Of the 10 *Patient global* domain items, 5 were general health items and 5 were DA items. There was evidence of local dependence between general health items and, separately, evidence of local dependence between DA items (Table 3). The residual principal components loadings also showed that all general health items loaded negatively, while all DA items loaded positively, on the first component (Table 3). Given this, two new domains were created: *General health* and *Disease activity*.

General health

For the five *General health* domain items, there were four items showing misfit, one item with DIF by sex, and local dependence between three items. It was decided to retain the other two items alongside one of these locally dependent items. Thus, three items (`R05`, `P01`, `C01`) were retained for the *General health* domain.

Disease activity

For the five *Disease activity* domain items, there were two items showing misfit, and all items were locally dependent on other items. There was a distinction in local dependence between the three items with a 6-month recall period and those with shorter recall periods. These three items were the only items among the 30 with a 6-month recall, so it was decided to retain the other two items (`PS1`, `A01`) in the *Disease activity* domain.

Pain

For the eight *Pain* domain items, there were five items showing misfit, and only one item was not locally dependent on another item. It was decided to retain three items [one with no local dependence (`F01`) and two with only minimal

evidence of local dependence between them (R04 and P07)] and one of the five locally dependent items. Four items (F01, R04, P07, Q05) were retained in the *Pain* domain that provided the best fit to the RMM.

Fatigue

The four *Fatigue* domain items demonstrated three items showing misfit, and DIF by age group and gender for one item. On retaining the three items without DIF, the analysis showed only a minor issue for item misfit, and therefore these three items (F03, PF1, RF1) were retained for the *Fatigue* domain.

Physical functioning

The two *Physical functioning* domain items (F02, F05) provided good fit to the RMM and were retained.

Stiffness

For the three items in the *Stiffness* domain, there was one item showing misfit, all items had disordered thresholds, and one item displayed DIF by earlier and accompanying diseases. There were two duration items, one of which had entirely illogical threshold ordering, and one intensity item. Therefore, the single intensity item (F04) was retained in the *Stiffness* domain.

Rasch measurement theory—validation dataset

Discrepancies

There was evidence of DIF by earlier and accompanying diseases for two items in the *General health* domain. For the *Pain* domain, the original item overdiscrimination issue remained, and another item also displayed misfit. A pair of items displayed local dependence, and unidimensionality could not be evidenced. For the *Fatigue* domain, there was evidence of item misfit and also DIF by highest educational qualification.

There were no discrepancies for the analyses of the *Tenderness and swelling*, *Disease activity* and *Physical functioning* domains, with no analysis for the *Stiffness* domain (only one item retained).

Domain-level testlets

None of the above discrepancies led to any need for changes to be made, therefore seven domain-level testlets representing the *Tenderness and swelling*, *General health*, *Disease activity*, *Pain*, *Fatigue*, *Physical functioning* and *Stiffness* domains were created (using the 18 retained items) and analysed. The *Fatigue* domain-level testlet had an extremely high positive fit residual (indicating underdiscrimination) and also displayed extremely large negative residual correlations with the *Tenderness and swelling*, *Disease activity*, *Pain*, *Physical functioning* and *Stiffness* domain-level testlets. This suggested that the *Fatigue* domain-level testlet did not measure the same construct as the other domain testlets (Fig. 2A), and it was, therefore, removed.

Analysis of the six remaining domain-level testlets provided a similar picture for the *General health* domain-level testlet: an extremely high positive fit residual (indicating underdiscrimination, Fig. 2B) and also extremely large negative residual correlations with all of the domain testlets. This suggested that the *General health* domain-level testlet did not measure the same construct as the other domain testlets, and it was, therefore, removed.

A final analysis of the five remaining domain-level testlets displayed issues, but none that required further change. There was item misfit for the *Disease activity* domain-level testlet with a large negative fit residual (indicating overdiscrimination) and a significant F-value. The *Physical functioning* domain-level testlet also had a large positive fit residual (indicating underdiscrimination). However, the item characteristic curves did not suggest any issues, so these were determined to be non-problematic. For the *Disease activity* domain-level testlet to exhibit overdiscrimination was logical, as it has the same wording as the construct of RA DA itself. The *Physical functioning* domain-level testlet is more of a functional status than a symptom status, so may underdiscriminate in comparison with the other domain-level testlets. There was evidence of local dependence between the *Disease activity* and *Tenderness and swelling* domain-level testlets and the *Physical functioning* and *Stiffness* domain-level testlets. Both of these combinations have conceptual sense in that RA DA inevitably causes tenderness and swelling, and greater levels of stiffness create issues with physical functioning. The *Pain* and *Physical functioning* domain-level testlets displayed DIF by age group, though this DIF was not evident graphically for the *Pain* domain-level testlet. For the *Physical functioning* domain-level testlet, it was logical that those participants aged 75 and over were at higher levels across the continuum in comparison with the other two age group categories. Also, unidimensionality could not be proven.

The 12 items therefore retained across the *Pain*, *Disease activity*, *Tenderness and swelling*, *Physical functioning* and *Stiffness* domains have their item codes highlighted in green in Supplementary Data S2, available at *Rheumatology* online.

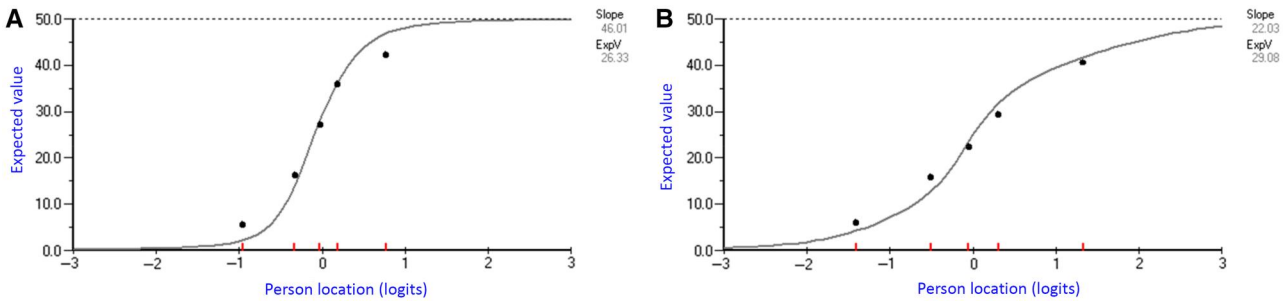
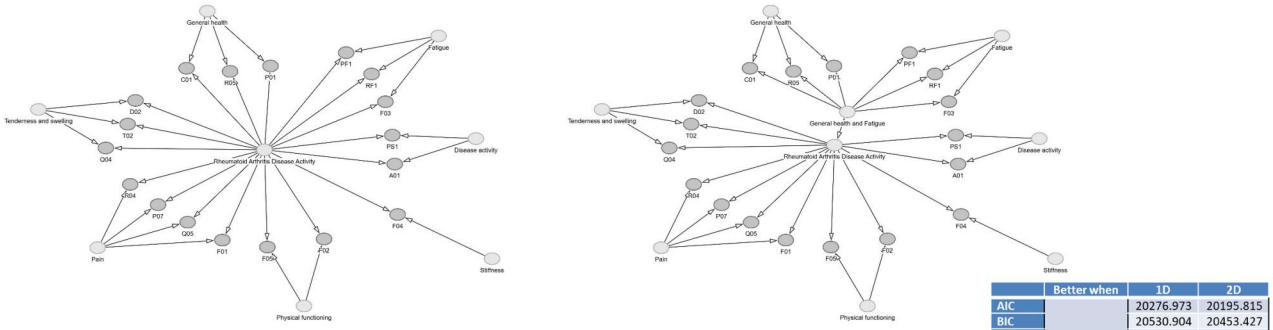


Figure 2. Item characteristic curves for the *Fatigue* domain-level testlet (from the analysis of all seven domain-level testlets, **A**) and for the *Patient global* domain-level testlet (from the analysis of six domain-level testlets minus *Fatigue,* **B**). The observed data (dots) should follow the ogive hypothesized by the Rasch measurement model. The observed data patterns here are flatter than the hypothesized ogive, indicating underdiscrimination

1-dimensional bifactor model (all items linked to construct of RA DA)

2-dimensional bifactor model (*Disease activity*, *Stiffness*, *Physical functioning*, *Pain* and *Tenderness and swelling* domain items linked to construct of RA DA; *General health* and *Fatigue* domain items linked to a separate construct to the construct of RA DA)



	Better when	1D	2D
AIC	Small	20276.973	20195.815
BIC		20530.904	20453.427
X^2		526.824	466.841
X^2/DF		4.390	3.923
RMSEA		0.108	0.100
SRMR		0.055	0.040
CFI	Large	0.929	0.940
TLI		0.910	0.922

Figure 3. Diagrammatical representations of the 1D and 2D bifactor models assessed by confirmatory factor analysis and results from these models. AIC: Akaike information criterion; BIC: Bayesian information criterion; DA: disease activity; DF: degrees of freedom; RMSEA: root mean square error of approximation; SRMR: standardized root mean square residual; CFI: comparative fit index; TLI: Tucker–Lewis index

Confirmatory factor analysis—validation dataset

CFA was used to assess and compare a 1D bifactor model and a 2D bifactor model, with a hypothesis that the 2D bifactor model would produce better summary statistics, as it better represented the model created through RMT analyses. This hypothesis was confirmed, as all summary values were better for the 2D bifactor model (Fig. 3).

Legacy patient-reported outcome measures—original dataset

For all legacy PROMs, the median of those having a flare was greater than the median of those not having a flare and, when compared through a Mann–Whitney *U* test, produced $P < 0.001$. (Supplementary Table S1, available at *Rheumatology* online). Spearman's ρ correlation coefficients were generally very high ($\rho \geq 0.833$ for RA DA PROMs) (Supplementary Table S2, available at *Rheumatology* online). Except for the PDAS2 variations, α was $\geq$0.802 across the PROMs (Supplementary Table S3, available at *Rheumatology* online). Details of the discretized visual analogue scale (VAS) items is shown in Supplementary Table S4, available at *Rheumatology* online. The CFA results show that only RADAI5, RADAI-SF and RA-FQ could evidence structural validity (Supplementary Table S5, available at *Rheumatology* online). RADAI5, RADAI, RADAI-SF, PDAS2, PRO-CLARA, GAS, PAS, PAS-II, RAPID3, RAPID4, PROM-score, RADAI-F5 and FLARE-RA Old did not fit the RMM (Supplementary Table S6, available at *Rheumatology* online), and all had misfitting items. Local dependence, disordered thresholds and DIF were issues across the majority of legacy PROMs. Unidimensionality could only be evidenced for PROM-score, RADAI-F5, FLARE-RA Arthritis, FLARE-RA General Symptoms and RA-FQ. The Person Separation Index was high for all PROMs, suggesting good levels of internal consistency. The measurement properties of the legacy PROMs are summarized in Supplementary Fig. S1, available at *Rheumatology* online.

Discussion

We undertook a cross-sectional study in plwRA to determine which items can form an item pool for measuring the construct of RA DA, and to examine the measurement properties of legacy RA DA PROMs and other relevant PROMs.

In analysing the initial domains under RMT, *General health* and *Disease activity* were found to be separate domains within the *Patient global* domain. By analysing domain-level testlets, it was found that 12 items across the *Pain*, *Disease activity*, *Tenderness and swelling*, *Physical functioning* and *Stiffness* domains can be used to form an item pool for a new PROM for measuring the construct of RA DA. *Fatigue* and *General health* domain items were shown through RMT analyses to measure a separate construct to the construct of RA DA.

Additionally, while all legacy PROMs had good evidence for internal consistency and hypothesis testing for construct validity, and many had evidence for structural validity from CFA, no legacy PROMs could fully evidence fit to the RMM.

The strength of this study is the novel and detailed strategy for analyses for the construct of RA DA. This was the first use of cross-validation (testing across two datasets) and RMT analyses for such items. This was the first use of CFA to complement RMT analyses, and the first use of bifactor models within CFA to confirm such an item structure. Equally, this was also the first time that RMT analyses have been applied to assess the measurement properties of legacy PROMs. There was also an adequate sample size to obtain reliable estimates through RMT analyses.

Patient and public involvement

J.D. and S.C., both plwRA, co-developed the participant information sheets, consent forms, and questionnaires. The National Rheumatoid Arthritis Society (NRAS) organized a focus group of 15 plwRA to discuss this research ahead of application.

Limitations

The data collected were from a small, densely populated area of South Wales, with an assumption that participants were able to understand the English language used in study documents and data collection forms. Collecting data from one geographical area meant that it was not possible to undertake simultaneous external validation with data from another area.

The paper questionnaire was very long, at 18 pages: this and other factors contributed to only 15.5% providing a response to all demographic questions and legacy PROM items. These questionnaires were also sent out at varying stages of the lockdowns enforced in Wales as a result of the COVID-19 pandemic. This may have discouraged potential participants from responding to the questionnaire, and possibly in different ways across distinct demographic groups. Further detail is available in Supplementary Data S4, available at *Rheumatology* online.

Future research

The next step is to undertake cognitive interviews with plwRA to assess the content validity measurement property. This will determine whether plwRA believe these items have relevance, comprehensiveness and comprehensibility in measuring the construct of RA DA. The 12 items have different recall periods, response formats and anchor wordings, so it will be important to explore preferences around these.

If this can be evidenced, then the item pool can be used to develop a computer adaptive test (CAT) or electronic PROM. However, there are only 12 items in the item pool, so the CAT will only provide a marginal burden reduction for plwRA, as a minimum of five items must be asked to cover all domains.

Supplementary material

Supplementary material is available at *Rheumatology* online.

Data availability

Data can be made available on request to the Centre for Trials Research https://www.cardiff.ac.uk/centre-for-trials-research/collaborate-with-us/data-requests.

Funding

T.P. was supported by a National Institute of Health Research Doctoral Fellowship, funded by the Welsh Government through Health and Care Research Wales (NIHR-FS-19).

Disclosure statement: E.C. has received research grants from Bio-Cancer, Biogen, Pfizer and Sanofi, and honoraria from / served as a member of speakers' bureaus for Abbvie, Bio-Cancer, Biocon, Biogen, Eli Lilly, Fresenius Kai, Galapagos, Janssen, Pfizer, Sanofi, UCB and Viatris. E.C. also receives a stipend as editor in chief of *Rheumatology (Oxford)*. N.M. has received consultancy fees from Novartis and Roche. N. M. has received other financial support from Abbvie, Amgen and UCB.

Acknowledgements

We would like to thank Kerry Nyland and Liz Griebel from the Research and Development Delivery Team in Cardiff and Vale UHB, Paula Phillips in the Rheumatology Department at Swansea Bay UHB, Keri Turner at the Research and Development Delivery Team, Ceril Rhys-Dillon and Catrin Margaret Jones in the Rheumatology Department in Cwm Taf Morgannwg UHB, and Anna Roynon the Research and Development Delivery Team in Aneurin Bevan UHB. We would like to thank Marcin Bargiel and Terri Kitson at the Centre for Trials Research for building the study database and for helping with the data entry, respectively.

References

1. Hendrikx J, de Jonge MJ, Fransen J, Kievit W, van Riel PL. Systematic review of patient-reported outcome measures (PROMs) for assessing disease activity in rheumatoid arthritis. RMD Open 2016;2:e000202.
2. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD; the Mayo FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007; 10: S94-S105.
3. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes 2006;4:79.
4. Mokkink LB, de Vet HCW, Prinsen CAC *et al.* COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1171–9.
5. Prinsen CAC, Mokkink LB, Bouter LM *et al.* COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1147–57.
6. Terwee CB, Prinsen CAC, Chiarotto A *et al.* COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018;27:1159–70.
7. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960.
8. Christensen KB, Kreiner S, Mesbah M. Rasch models in health. London: ISTE & Hoboken, NJ: John Wiley & Sons, 2013.
9. Fischer GH, Molenaar IW. Rasch models: foundations, recent developments, and applications. New York: Springer-Verlag, 1995.
10. Pickles T, Macefield R, Aiyegbusi OL *et al.* Patient Reported Outcome Measures for Rheumatoid Arthritis Disease Activity: a systematic review following COSMIN guidelines. RMD Open. 2022;8:e002093.
11. von Elm E, Altman DG, Egger M *et al.*; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806–8.
12. Linacre JM. Sample size and item calibration [or person measure] stability. Rasch Measur Trans 1994;7:328.
13. Leeb BF, Haindl PM, Brezinschek HP, Nothnagl T, Rintelen B. RADAI-5 to monitor rheumatoid arthritis. Clin Exp Rheumatol 2014;32:S-55–8.
14. Leeb BF, Haindl PM, Maktari A, Nothnagl T, Rintelen B. Patient-centered rheumatoid arthritis disease activity assessment by a modified RADAI. J Rheumatol 2008;35:1294–9.
15. Leeb BF, Sautner J, Mai HT *et al.* A comparison of patient questionnaires and composite indexes in routine care of rheumatoid arthritis patients. Joint Bone Spine 2009;76:658–64.
16. Fransen J, Langenegger T, Michel BA, Stucki G. Feasibility and validity of the RADAI, a self-administered rheumatoid arthritis disease activity index. Rheumatology (Oxford) 2000;39:321–7.
17. Stucki G, Liang MH, Stucki S, Brühlmann P, Michel BA. A self-administered rheumatoid arthritis disease activity index (RADAI)

for epidemiologic research. Psychometric properties and correlation with parameters of disease activity. Arthritis Rheum 1995; 38:795–8.
18. Veehof MM, ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Psychometric properties of the Rheumatoid Arthritis Disease Activity Index (RADAI) in a cohort of consecutive Dutch patients with RA starting anti-tumour necrosis factor treatment. Ann Rheum Dis 2008;67:789–93.
19. Choy EH, Khoshaba B, Cooper D, MacGregor A, Scott DL. Development and validation of a patient-based disease activity score in rheumatoid arthritis that can be used in clinical trials and routine practice. Arthritis Rheum 2008;59:192–9.
20. Leung AM, Farewell D, Lau CS, Choy EH. Defining criteria for rheumatoid arthritis patient-derived disease activity score that correspond to Disease Activity Score 28 and Clinical Disease Activity Index based disease states and response criteria. Rheumatology (Oxford) 2016;55:1954–8.
21. Salaffi F, Migliore A, Scarpellini M *et al.* Psychometric properties of an index of three patient reported outcome (PRO) measures, termed the CLinical ARthritis Activity (PRO-CLARA) in patients with rheumatoid arthritis. The NEW INDICES study. Clin Exp Rheumatol 2010;28:186–200.
22. Harrington JT. The uses of disease activity scoring and the physician global assessment of disease activity for managing rheumatoid arthritis in rheumatology practice. J Rheumatol 2009;36:925–9.
23. Wolfe F, Michaud K, Pincus T. A composite disease activity scale for clinical practice, observational studies, and clinical trials: the patient activity scale (PAS/PAS-II). J Rheumatol 2005;32:2410–5.
24. Pincus T, Bergman MJ, Yazici Y *et al.* An index of only patient-reported outcome measures, routine assessment of patient index data 3 (RAPID3), in two abatacept clinical trials: similar results to disease activity score (DAS28) and other RAPID indices that include physician-reported measures. Rheumatology (Oxford) 2008; 47:345–9.
25. Pincus T, Swearingen CJ, Bergman MJ *et al.* RAPID3 (Routine Assessment of Patient Index Data) on an MDHAQ (Multidimensional Health Assessment Questionnaire): agreement with DAS28 (Disease Activity Score) and CDAI (Clinical Disease Activity Index) activity categories, scored in five versus more than ninety seconds. Arthritis Care Res (Hoboken) 2010;62:181–9.
26. Bartlett SJ, Barbic SP, Bykerk VP *et al.* Content and construct validity, reliability, and responsiveness of the rheumatoid arthritis flare questionnaire: OMERACT 2016 workshop report. J Rheumatol 2017;44:1536–43.
27. Bykerk VP, Bingham CO, Choy EH *et al.* Identifying flares in rheumatoid arthritis: reliability and construct validation of the OMERACT RA flare core domain set. RMD Open 2016;2:e000225.
28. Berthelot J-M, De Bandt M, Morel J *et al.*; STPR Group of French Society of Rheumatology. A tool to identify recent or present rheumatoid arthritis flare from both patient and physician perspectives: the 'FLARE' instrument. Ann Rheum Dis 2012;71:1110–6.
29. de Thurah A, Maribo T, Stengaard-Pedersen K. Patient self-assessment of flare in rheumatoid arthritis: criterion and concurrent validity of the Flare instrument. Clin Rheumatol 2016;35:467–71.
30. Fautrel B, Alten R, Kirkham B *et al.* Call for action: how to improve use of patient-reported outcomes to guide clinical decision making in rheumatoid arthritis. Rheumatol Int 2018;38:935–47.
31. Maribo T, de Thurah A, Stengaard-Pedersen K. Patient-self assessment of flare in rheumatoid arthritis: translation and reliability of the Flare instrument. Clin Rheumatol 2016;35:1053–8.
32. Mason JH, Anderson JJ, Meenan RF *et al.* The rapid assessment of disease activity in rheumatology (radar) questionnaire. Validity and sensitivity to change of a patient self-report measure of joint count and clinical status. Arthritis Rheum 1992;35:156–62.
33. Mason JH, Meenan RF, Anderson JJ. Do self-reported arthritis symptom (RADAR) and health status (AIMS2) data provide duplicative or complementary information? Arthritis Rheum 1992; 5:163–72.
34. Hendrikx J, Fransen J, van Riel PL. Monitoring rheumatoid arthritis using an algorithm based on patient-reported outcome measures: a first step towards personalised healthcare. RMD Open 2015;1:e000114.
35. Hoque A, Gallagher K, McEntegart A *et al.* Measuring Inflammatory Foot Disease in Rheumatoid Arthritis: development and Validation of the Rheumatoid Arthritis Foot Disease Activity Index-5. Arthritis Care Res (Hoboken) 2021;73:1290–9.
36. Thorborg K, Holmich P, Christensen R, Petersen J, Roos EM. The Copenhagen Hip and Groin Outcome Score (HAGOS): development and validation according to the COSMIN checklist. Br J Sports Med 2011;45:478–91.
37. Boers M, Tugwell P, Felson DT *et al.* World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol Suppl 1994; 41:86–9.
38. Kirwan JR, Minnock P, Adebajo A *et al.* Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. J Rheumatol 2007;34:1174–7.
39. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Philos Mag J Sci 1901; 2:559–72.
40. Kreiner S. Validity and objectivity: reflections on the role and nature of Rasch models. Nord Psychol 2007;59:268–98.
41. Newby VA, Conner GR, Grant CP, Bunderson CV. The Rasch model and additive conjoint measurement. J Appl Meas 2009; 10:348–54.
42. Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 1969;34:183–202.
43. Kline RB. Principles and practice of structural equation modelling, 3rd ed. New York, NY: Guilford Press, 2011.
44. Muthen LK, Muthen B. Mplus version 8 user's guide. Los Angeles, CA: Muthen & Muthen, 2017. https://www.statmodel.com/ugexcerpts.shtml (3 February 2021, date last accessed).
45. Reise SP. The rediscovery of bifactor measurement models. Multivariate Behav Res 2012;47:667–96.
46. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18:50–60.
47. Hodges JL, Lehmann EL. Estimates of location based on rank tests. Ann Math Stat 1963;34:598–611.
48. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72–101.
49. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

Rheumatology, 2025, 00, 1–11
https://doi.org/10.1093/rheumatology/keaf189
Original Article