RHEUMATOLOGY

Original article

doi:10.1093/rheumatology/key154

Rasch measurement in rheumatoid arthritis: deriving psychometrically optimal measures from the Rasch Everyday Activity Limitation item bank

Martijn A. H. Oude Voshaar[1], Peter ten Klooster[1], Harald E. Vonkeman[1,2] and Mart. A. F. J. van de Laar[2]

Abstract

Objective. Recently we developed the Rasch Everyday Activity Limitations (REAL) generic item bank for measuring physical function. In this study we evaluate the REAL item bank in 209 RA patients and demonstrate how computerized adaptive testing (CAT) and Optimal Test Assembly methods can be used to derive measures from the REAL item bank with superior measurement performance compared with the HAQ Disability Index (HAQ-DI).

Methods. Structural validity of the item bank was assessed using confirmatory factor analysis. The validity of the REAL score metric in RA was evaluated by examining differential item functioning against the general population calibration sample. Besides the REAL-CAT, a 6-item short form (REAL-6) was developed using Optimal Test Assembly that was optimized with respect to common disability levels in RA, content and reading ease. Measurement precision of the different instruments was examined using item response theory methods. Construct validity was evaluated by testing hypothesized correlations with external measures.

Results. Good model-data fit was observed for a one-dimensional model and only two items showed differential item functioning of substantial magnitude. The REAL-CAT had superior measurement precision compared with HAQ-DI and REAL-6. REAL-6 outperformed HAQ-DI across all but the very lowest level of physical function. All three instruments demonstrated good construct validity (>75% of hypotheses affirmed) and only HAQ-DI had a ceiling effect (23.9%).

Conclusion. This study supports the validity of the REAL item bank and illustrates the potential of CAT and OTA applications based on the REAL item bank for assessing physical function in RA.

Key words: item response theory, Rasch, psychometric, validity, reliability, physical function, computerized adaptive test, optimal test assembly, activity limitations

CLINICAL SCIENCE

Rheumatology key messages

- The Rasch Everyday Activity Limitations item bank has favourable psychometric properties in RA.
- Computerized adaptive testing and targeted short forms yielded more precise scores than HAQ.
- Computerized adaptive testing and targeted short forms had lower ceiling/floor effects than HAQ.

Introduction

RA is an immune-mediated inflammatory disease that primarily affects the synovial joints. It is well known that pain, fatigue and joint damage associated with RA may impair patients' ability to perform everyday activities. Physical function is therefore a key outcome in this field [1].

[1]Arthritis Center Twente, Department of Psychology, Health and Technology, University of Twente, Enschede and [2]Arthritis Center Twente, Department of Rheumatology and Clinical Immunology, Medisch Spectrum Twente, Enschede, The Netherlands

Submitted 4 December 2017; revised version accepted 26 April 2018

Correspondence to: Martijn A. H. Oude Voshaar, Department of Psychology, Health and Technology, University of Twente, PO BOX 50 000, 7500 KA Enschede, The Netherlands.
E-mail: A.H.oudevoshaar@utwente.nl

The Rasch model is increasingly used in the development or evaluation of instruments for measuring self-reported physical function in various settings. It is part of the larger family of parametric item response theory (IRT) models, which provides a theoretically well-founded framework in which parameters representing, respectively, the persons and items are mapped on a common underlying measurement continuum [2]. In RA, Rasch modelling has thus far primarily been used to help develop or study the measurement properties of fixed-length physical function scales [3–7].

IRT scaling procedures can also be used to collect the measurement characteristics of a large number of items that measure the same construct in a so-called item bank. Once it exists, an item bank can be used to develop new

measures by drawing items from the bank. Because IRT models adjust scores for item characteristics, any number and combination of items can be selected, yet scores will still be comparable to other measures derived from the item bank. Furthermore, measures derived from an item bank will have predictable psychometric characteristics. This is because the contribution of individual items to the precision of scores for different levels of the measured attribute is described for all items in the item bank by item information functions [8].

Two different methods have been described in the literature that help researchers to develop psychometrically optimal measures by automating the item selection process. Computerized adaptive testing (CAT) is the best-known procedure in which scores of individual respondents are optimized in real-time by having a computer algorithm administer subsequent items that are most appropriate for the estimated level of the respondent based on item responses up until that point [9]. Optimal Test Assembly (OTA) is an alternative approach for deriving fixed-length questionnaires. It differs from CAT in that the most suitable items are selected for a range of trait levels that are considered relevant *a priori* [10, 11].

Procedures such as CAT and OTA could be used to derive patient-reported physical function measures from an item bank with improved precision and associated increased power to detect treatment effects in studies with RA patients [12]. Such measures could even be shorter than currently used questionnaires. OTA methods can also easily be used to optimize non-statistical characteristics of physical function measures by imposing constraints on the item selection process. Finally, by deriving CAT or OTA applications from a calibrated item bank, comparability across studies that have used different forms could be maintained, eliminating the need to use the same physical function questionnaire in each study, as is effectively the case now with the HAQ Disability Index (HAQ-DI).

Recently, we developed and calibrated the generic Rasch Everyday Activity Limitations (REAL) item bank for measuring physical function in a large sample representative of the general population [13]. In the present study we evaluate its psychometric properties in a sample of RA patients and illustrate how CAT and OTA applications can be used to derive measures with superior measurement performance compared with the HAQ-DI.

Methods

Patients and study procedures

Two hundred and seventy-seven consecutive patients with RA from the Rheumatology clinic at Medisch Spectrum Twente who are enrolled in the Dutch rheumatoid arthritis measurement study were invited to participate in this study via the web-portal they routinely use to fill out questionnaires in preparation for their scheduled clinic visits. Two hundred and nine (75%) patients agreed to participate. According to the Dutch Medical Research Involving Human Subjects Act, the study did not need approval of an ethical review board because participants were not subjected to medical procedures nor required to follow rules of behaviour that potentially infringe on their physical or psychological integrity. Nevertheless, each respondent provided written consent to be included, and data collection abided by the Dutch protection of personal data act. Upon digitally providing informed consent they filled out the full REAL item bank and the HAQ-DI outlined below. Additionally, patients completed the Medical Outcomes Study Short Form 36 (SF-36) health survey v2.0, which is a 36-item health index that assesses health in the eight domains of health-related quality of life. The SF-36 has been extensively validated in RA [14]. Visual analogue scales were used to assess pain, fatigue and overall well-being. Disease activity was assessed using the DAS28 [15]. Assessments of joint activity and laboratory measures were obtained during the clinic visit.

Physical function measures

REAL is a recently developed item bank, calibrated using the Rasch partial credit model [16]. Compared with more elaborate IRT models, the partial credit model has the advantage that fewer parameters need to be estimated and no assumptions need to be made about the distribution of scores when estimating the item parameters. It contains 41 items that assess disability in everyday activities that people engage in as part of their daily domestic routine [13]. The majority (70%) of the items were generated directly from activities that patients reported as most bothersome in a previous survey study. It was developed to better match disability levels of outpatient populations compared with instruments that focus on basic activities of daily living, such as the HAQ-DI, by including items that assess health concepts related to the domestic life chapter of the International Classification of Functioning Disability and Health. REAL items have a 5-point response scale ranging from 0 (without any difficulty) to 4 (unable to do). REAL-derived physical function estimates are expressed on a metric with higher scores denoting better functioning. Scores in the general Dutch population were approximately normally distributed with a mean (S.D.) score of 50.62 (17.54). A previous study demonstrated superior measurement performance of REAL-derived CAT compared with the SF-36 physical functioning scale in the Dutch general population [13].

The HAQ-DI assesses disability in eight categories of basic daily activities (no. of items): Dressing (two), Eating (three), Grip (three), Hygiene (three), Reaching (two), Rising (two), Walking (two) and Activities (three), using a rating scale ranging from 0 (without any difficulty) to 3 (unable to do) [17]. Respondents are also asked if they use help of others or aids to perform the activities. Category scores are obtained by taking the item with the highest level of disability within each category. In case a respondent reports using aids or help of others for any activity within a category, the category score is set to 2 if the unadjusted category score was <2. The HAQ-DI score is obtained by taking the mean of these category scores. The HAQ-DI is probably the most used physical function instrument in RA and previous studies

have supported its reliability, construct validity and responsiveness [18].

Information maximization

The REAL item bank was calibrated using the partial credit model. In this model item parameters and persons are represented as fixed positions on an underlying scale. Item parameters mark the location on the scale where consecutive response options are equally likely to be chosen. The amount of information provided by one item depends on how well targeted the item is to the trait level of the respondent. The standard error of estimation (SEE), which is a measure of the precision with which a score has been estimated, is inversely related to information (i.e. $SEE = \frac{1}{\sqrt{I}}$). Both OTA and CAT seek to draw items from the item bank that maximize information at relevant levels of the measured attribute. For CAT, information is maximized at a single point on the latent variable, corresponding to the current estimated physical function level of an individual respondent. For OTA, information is maximized for an *a priori* defined range of physical function levels, intended to reflect the distribution of physical function scores in the target population.

OTA-based short form development

We used the mixed integer programming method proposed by Van der Linden [10] to derive a short form that includes as few items as possible, while still meeting a number of specifications. We wanted to include a proportional number of items reflecting the three main International Classification of Functioning, Disability and Health-based content areas included in the REAL item bank: acquisition of necessities, taking care of plants or animals and assisting others, and caring for household objects and household tasks. We also wanted the short form to include items that were as easy to read as possible, and we wanted the short form to yield precise scores (i.e. reliability >0.90) across the REAL score levels ranging from 20 to 60. These levels were chosen since we observed in this dataset, as well as other datasets, that about 90% of RA patients had a REAL score between 20 and 60. These requirements can be framed as a combinatorial optimization problem, where the objective function is to minimize the number of items included in the short form, subject to the constraints that the percentage of items belonging to each of the three International Classification of Functioning, Disability and Health-based content areas should not exceed 34% and the mean Flesch-Kincaid Reading Ease score [19] should be >60. The SEE should be $\geqslant 0.5$, which is roughly equivalent to classical reliability of 0.90 [20], at each of the REAL score levels 20, 40 and 60.

The algorithm was programmed in Excel, using the Solver add-in.

CAT simulation

A CAT simulation based on the REAL item bank data was performed in a Bayesian framework, with the maximum posterior weighted information procedure for interim item selection and the expected *a posteriori* method for interim trait level estimation, in both cases with a standard normal prior. Final estimates and their standard errors were obtained using the maximum likelihood procedure. Administration of items was stopped in case the SEE was <0.44, which is roughly equivalent to classical reliability of 0.95 [21], or if 20 items were administered, whichever occurred first.

Statistical analysis

Analysis of IRT assumptions and model fit of the REAL item bank in RA

We evaluated the assumption of unidimensionality in the RA population using confirmatory factor analysis on polychoric correlations using MPLUS. Fit of a unidimensional model was evaluated using the fit indices provided by MPLUS. Values for Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) >0.95 and Root Mean Square Error of Approximation (RMSEA) <0.08 are generally considered to reflect good fit [21].

The validity of the general population-based REAL score metric in RA was evaluated using Lagrange multiplier (LM) statistics for differential item functioning (DIF) and associated effect size (ES) statistics [22]. Items were considered to exhibit DIF of unacceptable magnitude in case the LM test was statistically significant ($P \leqslant 0.05$) and the average model-predicted scores deviated $>5\%$ from their expectation under the model (i.e. ES $>5\%$). IRT analyses were performed using the Lexter software package [23].

Evaluation of measurement properties of the REAL-CAT and OTA applications in RA

The precision with which IRT scores of patients with different levels of physical function could be estimated with the different instruments was evaluated by comparing the SEE of the various instruments across the REAL score continuum. For the REAL short form and HAQ-DI, SEE were derived from the appropriate item information functions, for CAT these were obtained from the output of the analysis. Convergent validity of the different physical function instruments was evaluated by testing hypotheses about the correlation of scores with related measures that were defined before data collection was started, which is the recommended best practice for assessing construct validity [24]. To achieve this, we used the exact hypotheses that were used in a previously published systematic review on measurement properties of physical function scales in RA [18]. In that review it was hypothesized that strong correlations ($r > 0.60$) would be found with patient-reported outcome measures (PROMs) of physical aspects of health, and moderate correlations ($0.30 < r < 0.60$) with PROMs of non-physical aspects of health and clinical outcome measures. Furthermore, it was hypothesized that the strongest correlations would be observed with other measures of physical function, followed by PROMs of physical aspects of health, PROMs of non-physical aspects and clinical measures, respectively. To test the second type of hypotheses, the correlations of all instruments assessing a domain were averaged, and the rank order of these correlations was

examined. Instruments were considered to have good construct validity in case >75% of the hypotheses would not be rejected [25]. Both floor and ceiling effects were compared between instruments. These were defined as the percentage of patients always selecting the response option reflecting the worst and best level of physical function, respectively.

Results

In total, 209 patients completed the REAL item bank and other patient-reported outcomes. Baseline characteristics are presented in Table 1. Most patients were female and generally had well controlled disease, with low disability according to HAQ-DI. However, according to REAL scores physical function levels were below the general population average. The number of missing responses was very low (max = 2) for the REAL items. For 12 patients, no HAQ-DI could be calculated.

Analysis of Rasch model fit and assumptions

The fit of a unidimensional measurement model was good according to the fit indices (CFI = 0.98, TLI = 0.98, RMSEA = 0.08). All factor loadings were high and, in accordance with the Rasch model assumption of equal discrimination of items, similar for different items [mean (S.D.) = 0.90 (0.03)]. DIF of unacceptable magnitude was found only for items 4 (LM = 53.66, $P < 0.01$, ES = −5.2%) and 26 (LM = 30.43, $P < 0.01$, ES = −5.2%). In both cases, item scores of RA patients were lower than expected by the model. To remedy this, these items were assigned disease-specific item parameters in subsequent analyses. Fit of the re-specified model was good, with none of the items meeting the criteria for substantial lack of fit.

Table 1 Baseline characteristics

Variable	Patients with RA (*n* = 209)
Age, mean (S.D.), years	57.5 (11.78)
Female, *n* (%)	125 (59.8)
Disease duration, mean (S.D.), years	9.50 (7.95)
REAL score, mean (S.D.)	38.17 (18.48)
HAQ-DI, mean (S.D.)	0.69 (0.67)
VAS pain, mean (S.D.)	27.99 (25.96)
VAS fatigue, mean (S.D.)	37.43 (30.53)
DAS28, mean (S.D.)	2.19 (1.08)

REAL: Rasch Everyday Activity Limitations; HAQ-DI: HAQ Disability Index; VAS: visual analogue scale.

OTA based short form generation and analysis of feasibility

A 6-item short form, which will be referred to as REAL-6, was generated that met all the pre-specifications. The included items are presented in Table 2. The mean (S.D.) Flesch-Kincaid Reading Ease score for the included items was 62.01 (8.80). Mean (S.D.) reading ease for the total item bank was 54.61 (15.56). Marginal reliability of the 6-item short form was 0.93.

Measurement performance

Both REAL-derived instruments outperformed the HAQ in terms of measurement range, with the range of REAL scores for which scores were reliable (i.e. reliability coefficient >0.70) extending from 0 to 70 for HAQ and 0–90 for REAL-6 and REAL-CAT (Fig. 1). CAT scores consistently reached the pre-specified level of precision over the range of scores from 0 to 70. Precision of CAT decreased slightly for higher scores and the maximum number of 20 items were administered for each respondent at these score levels. This happened because all the item threshold parameters have values <80 on the REAL score continuum. However, even at these high levels of physical function, scores were still very precise (classical reliability >0.90).

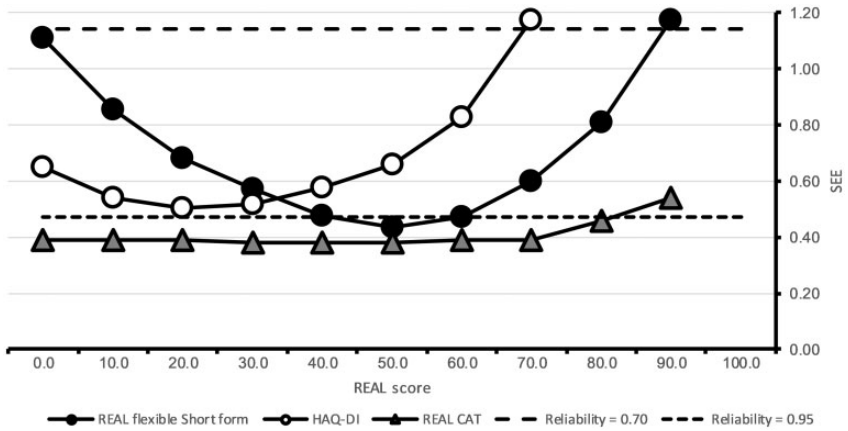
The results of the analysis of construct validity are presented in Tables 3 and 4, together with information about the hypothesized magnitude of correlations. Since the SF-36 also has a physical functioning scale (the PF-10), it was also added to the columns of the tables. The pattern of correlations with other measures was very similar for the HAQ-DI, SF-36 PF-10 and both REAL-derived instruments, suggesting the instruments assess a similar construct. The percentage of affirmed hypotheses was highest for the REAL-CAT (85.5%), followed by the

Table 2 Items included in REAL-6 short form

Item	ICF code	Item description	Reading Ease	β_1	β_2	β_3	β_4
1	d6550	Heavy ironing	35.5	34	27	17	17
5	d65501	Making repairs around the house	60.7	52	40	31	32
14	d65505	Mowing the lawn with a hand mower	67.7	45	39	32	25
17	d65505	Walking a leashed dog	83.3	24	21	16	27
31	d6201	Carrying bags of fruit	71.0	41	22	20	16
32	d6201	Carrying groceries	53.9	35	20	16	8

REAL: Rasch Everyday Activity Limitations; ICF: International Classification of Functioning, Disability and Health; Reading Ease: Flesch-Kincaid Readings Ease score;β: item threshold parameter.

Fig. 1 Measurement precision of different instruments across physical function continuum



REAL: Rasch Everyday Activity Limitations; HAQ-DI: HAQ Disability Index; SEE: standard error of estimation.

Table 3 Construct validity hypotheses testing based on Pearson correlations between physical function scores and related and unrelated measures

Instrument	Expected correlation	Observed correlations			
		CAT	REAL-10	HAQ-DI	SF-36 physical functioning
CAT	>0.60			0.75	0.73
REAL-10	>0.60			0.75	0.73
HAQ-DI	>0.60	0.75	0.84		0.82
VAS pain	>0.60	0.44	0.49	0.6	0.55
SF-36 bodily pain	>0.60	0.61	0.63	0.68	0.69
SF-36 physical functioning	>0.60	0.73	0.83	0.82	
SF-36 physical role	>0.60	0.63	0.69	0.7	0.66
SF-36 general health	0.30< r <0.60	0.57	0.63	0.57	0.59
SF-36 vitality	0.30< r <0.60	0.56	0.62	0.61	0.58
SF-36 social function	0.30< r <0.60	0.57	0.60	0.65	0.65
SF-36 role emotional	0.30< r <0.60	0.50	0.55	0.58	0.52
SF-36 mental health	0.30< r <0.60	0.34	0.38	0.46	0.36
VAS fatigue	0.30< r <0.60	0.56	0.59	0.54	0.52
VAS wellbeing	0.30< r <0.60	0.49	0.48	0.60	0.59
VAS disease activity	0.30< r <0.60	0.38	0.39	0.70	0.57
TJC 28	0.30< r <0.60	0.26	0.34	0.38	0.30
SJC 28	0.30< r <0.60	0.14	0.07	0.16	0.11
DAS28	0.30< r <0.60	0.43	0.45	0.54	0.52

CAT: computerized adaptive test; REAL: Rasch Everyday Activity Limitations; HAQ-DI: HAQ Disability Index; VAS: visual analogue scale; SF-36: Medical Outcomes Study Short Form 36; TJC: tender joint count; SJC: swollen joint count.

REAL-6 (83%). However, also for the HAQ-DI (76.2%) and SF-36 PF-10 (80.1%), the percentage of confirmed hypotheses was >75%, indicating good construct validity for all physical function instruments. The HAQ-DI was the only instrument with a substantial ceiling effect with 23.9% of patients scoring the best possible score. The percentage of patients at the ceiling was lowest for the REAL-CAT (4.8%). Floor effects were absent, with the percentages of patients at the floor <2.5% for each instrument.

Discussion

The present study supports the validity of the REAL item bank in patients with RA and illustrates how CAT and OTA can be used to derive measures from the REAL item bank that yield more precise scores compared with the HAQ-DI.

In the first part of this study we examined the psychometric properties of the total REAL item bank when applied in RA. The results of the confirmatory factor

Table 4 Construct validity hypotheses testing based on Pearson correlations between physical function scores and aggregated intra-domain scores

Instrument	Physical function	Physical PRO	Other PRO	Clinical measures
CAT	0.74	0.56	0.51	0.30
REAL-6	0.83	0.61	0.56	0.31
SF-36 physical functioning	0.76	0.63	0.54	0.37
HAQ-DI	0.77	0.65	0.57	0.44

CAT: computerized adaptive test; SF-36: Medical Outcomes Study Short Form 36; REAL: Rasch Everyday Activity Limitations; HAQ-DI: HAQ Disability Index; PRO: patient-reported outcome.

analysis supported the unidimensionality of the item bank in this patient population. Furthermore, subsequent DIF analysis demonstrated that, overall, patients with RA respond similarly to people from the Dutch general population to most of the items in the item bank, when the differences in disability levels between these populations are taken into account. These findings support the validity of REAL scores obtained using the adapted model and applications that use the REAL item parameters such as CAT and OTA in the RA population [26].

In the second part of this study, we demonstrated that measures derived from the REAL item bank assess a similar construct as the HAQ-DI and compared their measurement precision with this instrument. The REAL-CAT clearly outperformed the REAL-6 and the HAQ-DI considering that the specified measurement precision targets were consistently met for most of the measurement continuum. Superior measurement precision will translate into more efficient testing in clinical trial settings [12]. If a trial were to be powered on its ability to detect treatment effects on physical function, fewer patients would need to be included when using the REAL-CAT compared with a static fixed-length instrument. In the most common case, however, where physical function is a secondary outcome, studies would have more power to detect treatment effects on physical function if they employed the CAT derived from the REAL item bank. Furthermore, less than half the number of items were used to obtain REAL-CAT scores on average compared with the HAQ-DI. This is an important practical gain, because in most clinical research settings patients have to fill out a variety of questionnaires, which may detract from their motivation to stay engaged with the study.

However, CAT applications may not always be feasible in clinical practice or research settings, since scores need to be estimated using complicated scoring algorithms, in real time. In such cases, OTA provides a powerful alternative that still allows the two principle advantages of IRT-based measurement to be realized in practice. This was evidenced by our finding that the REAL-6 clearly outperformed the HAQ-DI in terms of efficiency and measurement precision, with only six items. The REAL-6 assessed a similar construct as the HAQ-DI according to the results of the analysis of construct validity, and it can be expected to perform better than the HAQ-DI based on the results of this study. Moreover, unlike the HAQ-DI it assesses the patient's ability to engage in everyday domestic activities, including important aspects of physical disability that have, in our opinion, been neglected in existing physical function questionnaires, such as physical disabilities relating to helping other people in the household and taking care of one's domestic responsibilities. In this sense, REAL-derived measures could complement existing fixed-length physical function questionnaires. Those interested in using the REAL-6 in their studies could average responses to its six questions to obtain a total score. In case of missing data or to compare outcomes with other studies that have used REAL-derived measures, the item parameters provided in the original paper can also be used to estimate REAL scores [13]. We would also encourage those interested to make their own selections of items to administer and consult the original publication for guidance on how to obtain REAL scores.

To the best of our knowledge, only one previous study has used OTA methods to derive short forms in rheumatology [11]. In that study, a procedure is proposed where short forms with different numbers of items are generated, with the objective to maximize the scale information curve. These short forms are then manually compared with the shortest possible short form that still meets a number of statistical performance requirements. In our study we propose a procedure for finding the shortest possible short form where the number of items is the objective function that needs to be minimized and the psychometric requirements are programmed as constraints, so that the results do not need to be manually checked.

Despite the encouraging findings in the present study, we were not yet able to study the responsiveness of the REAL measures due to the cross-sectional design of this study. Based on the results of previous simulation studies, which have shown that the increased precision of CAT applications translates to increased power, we expect that the REAL-CAT and to a lesser extent REAL-6 will perform well in detecting changes in physical functioning [12]. However, few studies have yet compared the sensitivity to change of actual CAT administrations with fixed-length questionnaires. Moreover, the small sample size was a limitation to the DIF analysis. It will also be important to study the measurement properties of REAL-derived measures using data that was specifically collected in their respective specified form, and not only in data from total item bank administrations. Also, most patients had well-controlled disease and relatively mild disability in the current study. Lower ceiling effects would have likely occurred in patients with active disease. Finally, content validity of the REAL-derived measures should be studied by having patients review the items.

In summary, the present study supports the psychometric properties of the REAL item bank applied in RA. We were also able to demonstrate considerable gains in

measurement precision using REAL-derived CAT and OTA measures compared with the HAQ-DI, while at the same time decreasing the average number of items that patients have to respond considerably.

Funding: No specific funding was received from any funding bodies in the public, commercial or not-for-profit sectors to carry out the work described in this manuscript.

Disclosure statement: The authors have declared no conflicts of interest.

References

1 Felson DT, Anderson JJ, Boers M *et al*. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729–40.

2 Fischer G, Molenaar I. Rasch Models: Foundations, Recent Developments, and Applications. 2012. https://books.google.com/books? hl=nl&lr=&id=AcrgBwAAQBAJ&oi=fnd&pg=PA3&dq=Rasch+Models+-+Foundations,+Recent+Developments,+a&ots=B2bFAU5aQg&sig=uiA-nCo03dghDsXN0waiyhgj6Vs (9 July 2017, date last accessed).

3 Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum 2007;57:723–9.

4 ten Klooster PM, Taal E, van de Laar MAFJ. Rasch analysis of the Dutch Health Assessment Questionnaire disability index and the Health Assessment Questionnaire II in patients with rheumatoid arthritis. Arthritis Rheum 2008;59:1721–8.

5 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 Rheum 2004;51:14–9.

6 Leung Y-Y, Png M-E, Conaghan P, Tennant A. A systematic literature review on the application of Rasch analysis in musculoskeletal disease – a special interest group report of OMERACT 11. J Rheumatol 2014;41:159–64.

7 Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum 2004;50:3296–305.

8 Samejima F. A use of the information function in tailored testing. Appl Psychol Meas 1977;1:233–47.

9 van der Linden WJ, Glas C. Elements of Adaptive Testing. New York: Springer, 2010.

10 van der Linden WJ. Linear Models for Optimal Test Design. 2006. https://books.google.nl/books?hl=nl&lr=&id=uAd1eYJniBQC&oi=fnd&pg=PA1&dq=Linear+Models+for+Optimal+Test+Design&ots=H0elYOFLx9&sig=ReLSfJc4Z0HA8AC1Omgqa57BlyM (27 May 2016, date last accessed).

11 Levis AW, Harel D, Kwakkenbos L *et al*. Using optimal test assembly methods for shortening patient-reported outcome measures: development and validation of the Cochin Hand Function Scale-6 - a Scleroderma Patient-centered Intervention Network (SPIN) Cohort Study. Arthritis Care Res (Hoboken) 2016;68:1704–13.

12 Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. Control Clin Trials 2003;24:390–410.

13 Oude Voshaar MAH, Ten Klooster PM, Vonkeman HE, van de Laar MAFJ. Measuring everyday functional competence using the Rasch assessment of everyday activity limitations (REAL) item bank. Qual Life Res 2017;26:2949–59.

14 ten Klooster PM, Vonkeman HE, Taal E *et al*. Performance of the Dutch SF-36 version 2 as a measure of health-related quality of life in patients with rheumatoid arthritis. Heal Qual Life Outcomes 2013;11: Article number 77.

15 Prevoo ML, van't Hof MA, van Kuper HH *et al*. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.

16 Masters GA. A rasch model for partial credit scoring. Psychometrika 1982;47:149–74.

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

18 Oude Voshaar MAH, ten Klooster PM, Taal E *et al*. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Heal Qual Life Outcomes 2011;9: Article number 99.

19 Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. 1975. http://www.dtic.mil/docs/citations/ADA006655 (17 July 2017, date last accessed).

20 Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. Qual Life Res 2007;16:109–19.

21 Yu C. Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes. 2002. https://pdfs.semanticscholar.org/7a22/ae22553f78582fc61c6cab4567d36998293b.pdf. (16 February 2017, date last accessed).

22 Glas CAW. Detection of differential item functioning using Lagrange multiplier tests. Stat Sin 1998;8:647–67.

23 Glas CAW. Preliminary Manual of the Software Program Lexter; 2017.

24 Mokkink LB, Terwee CB, Patrick DL *et al*. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539–49.

25 Terwee CB, Bot SDM, de Boer MR *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.

26 Hambleton R, Swaminathan H, Rogers H. Fundamentals of Item Response Theory. Newbury Park, CA: Sage, 1991.