ARTHRITIS & RHEUMATOLOGY
Vol. 66, No. 10, October 2014, pp 2900–2908
DOI 10.1002/art.38759

Relative Performance of Commonly Used Physical Function Questionnaires in Rheumatoid Arthritis and a Patient-Reported Outcomes Measurement Information System Computerized Adaptive Test

Martijn A. H. Oude Voshaar,[1] Peter M. ten Klooster,[1] Cees A. W. Glas,[2] Harald E. Vonkeman,[3] Eswar Krishnan,[4] and Mart A. F. J. van de Laar[3]

Objective. **To evaluate and compare the measurement precision and sensitivity to change of the Health Assessment Questionnaire disability index (HAQ DI), the Short Form 36 physical functioning scale (PF-10), and simulated Patient-Reported Outcomes Measurement Information System (PROMIS) physical function computer adaptive tests (CATs) with 5, 10, and 15 items, using item response theory–based simulation studies.**

Methods. **The measurement precision of the various physical function instruments was evaluated by calculating root mean square errors (RMSEs) between true physical function levels (latent physical function score) and estimated physical function levels. Measurement precision was evaluated at 9 levels of physical function, with 5,000 simulated response patterns per level. Sensitivity to change was evaluated by the ability of a simple statistical test to detect simulated change scores of small to moderate magnitude (standardized effect sizes 0.20, 0.35, and 0.50).**

Results. **RMSEs were smaller for the PROMIS physical function 15-item CAT (CAT-15) and CAT-10 than for the HAQ DI and PF-10 across all levels of the latent physical function scale. Only marginal improvement in performance was observed for the CAT-15 compared with the CAT-10, and the CAT-5 performed quite similarly to the HAQ DI and PF-10 across most levels of the latent physical function scale. Substantially improved sensitivity to change was observed for the CAT-10 compared with the HAQ DI and PF-10, particularly in detecting moderate effect sizes.**

Conclusion. **Clearly higher measurement precision was observed for the PROMIS CAT compared with the HAQ DI and PF-10. Higher reliability also translated into lower sample size requirements for detecting changes in clinical status.**

Rheumatoid arthritis (RA) is a chronic, systemic inflammatory disease characterized by progressive inflammation of connective tissue. Inflammation of synovial membranes damages the joint capsule and articular cartilage. This characteristic joint damage evolves slowly over the course of the disease, with an incremental effect on the ability of the patient to perform routine daily activities. Therefore, preservation of physical function is a key therapeutic goal in the long-term management of RA, and physical function has been an endorsed study end point since the first proposed core set of outcome measures in this field (1,2).

Most contemporary clinical and observational studies in RA use self-report questionnaires to assess physical function (3). However, the concurrent assessment of patient-reported physical function and the increasing number of other relevant patient-reported outcome domains place a considerable and increasing burden on both the patient and the administrator. Moreover, it has proven difficult to develop feasible fixed-length questionnaires that adequately measure the

[1]Martijn A. H. Oude Voshaar, MSc, Peter M. ten Klooster, PhD: Arthritis Centre Twente and University of Twente, Enschede, The Netherlands; [2]Cees A. W. Glas, PhD: University of Twente, Enschede, The Netherlands; [3]Harald E. Vonkeman, MD, PhD, Mart A. F. J. van de Laar, MD, PhD: Arthritis Centre Twente, University of Twente, and Medisch Spectrum Twente, Enschede, The Netherlands; [4]Eswar Krishnan, MD: Stanford University School of Medicine, Palo Alto, California.

Address correspondence to Martijn A. H. Oude Voshaar, MSc, Department of Psychology, Health, and Technology, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: A.H.OudeVoshaar@utwente.nl.

Submitted for publication April 8, 2014; accepted in revised form June 19, 2014.

variety of physical function levels that occur during various stages of RA. Consequently, floor and ceiling effects are common when standard outcome measures such as the Health Assessment Questionnaire disability index (HAQ DI) (4) are used, particularly in relatively well-functioning populations (5,6). This has led to increased interest in the flexibility of item response theory (IRT)–based computerized adaptive assessment of physical function (7–9).

In the IRT framework, item and person parameters are calibrated on a latent metric commonly referred to as θ. Individual items are described by item characteristic functions that provide the probability of a given response as a function of θ (10). After a set of items has been calibrated under an IRT model, maximum likelihood or Bayesian procedures can be used to estimate θ for future respondents within the same population, from any subset of items. Calibrated item banks can therefore be used to create more efficient measures by administering only the most relevant questions for specific research needs, either by manually selecting items or using computerized adaptive testing (CAT) algorithms, in which physical function estimates of individual patients are sequentially statistically optimized (11). This is achieved by capitalizing on the local definition of reliability in IRT, where so-called information functions describe the measurement precision of individual items at each of the different levels of physical function that are measured by the ensemble of the calibrated items.

Given the practical constraint that only a limited number of items can be administered, the most reliable physical function estimates are obtained if those questions are administered that have the highest measurement precision at the (estimated) level of physical function of the responding patient. Theoretically, this means that more precise and optimally efficient estimates of physical function may be obtained with CATs compared with classic instruments in which the same number of untailored items is administered to each patient, provided that the CAT algorithm can select items from a sufficiently suitable item bank.

Thus far, however, no studies have directly evaluated whether these theoretical advantages translate into meaningful practical gains in measurement precision and efficiency for assessing physical function in RA. Recently (12), we concurrently calibrated a Dutch-Flemish version of the Patient-Reported Outcomes Measurement Information System (PROMIS) physical function item bank and both the HAQ DI (4) and Short Form 36 (SF-36) (13) physical functioning scale (PF-10) for use in RA.

The objective of the current study was to evaluate and compare the measurement precision of a PROMIS physical function CAT with 5, 10, and 15 items with the measurement precision of the HAQ DI and the PF-10, which are currently the most frequently used tools in this field for assessing physical function. Furthermore, we evaluated whether observed differences in measurement precision would also yield more power to detect changes in functional status. Both objectives were evaluated by means of simulation studies.

PATIENTS AND METHODS

Measures. *SF-36 physical functioning scale.* The PF-10 is one of the 8 scales of the SF-36 health survey (13) and consists of 10 items measuring perceived current limitations in a variety of physical activities on a 3-point response scale from 1 (yes, limited a lot) to 3 (no, not limited at all). Scores for the PF-10 items are summed and linearly transformed to range from 0 to 100, with higher scores indicating more favorable levels of physical functioning (13).

HAQ DI. The HAQ DI contains 20 items measuring physical disabilities over the past week in 8 categories of daily living: dressing and grooming, rising, eating, walking, hygiene, reach, grip, and activities. Each item is scored on a 4-point rating scale from 0 (without any difficulty) to 3 (unable to do). Eight category scores are obtained by ranking the worst item score in each of the respective categories, and a total HAQ DI score is obtained by averaging the 8 category scores (4).

PROMIS physical function item bank. The PROMIS physical function item bank measures an individual's self-reported, current capability to carry out activities that require physical actions, ranging from self-care (activities of daily living) to more complex activities that require a combination of skills, often within a social context. The final calibrated item bank contains 121 questions assessing the functioning of the upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living, such as running errands. Each item is scored on a 5-point rating scale, with higher scores indicating better functioning. The Dutch-Flemish translation of the item bank was developed according to the universal PROMIS translation approach, which included extensive forward-back translation procedures, expert reviews, and cognitive debriefing interviews among Dutch and Flemish participants (8,14).

Item bank calibration. The data that were used to calibrate the item bank were collected from 690 patients with RA who were participants in the Dutch Rheumatoid Arthritis Monitoring registry. Consecutive patients were invited to participate in the study when they logged on to their patient portals. The generalized partial credit model (GPCM) was used to concurrently calibrate the HAQ DI, PF-10, and PROMIS physical function item bank in a single IRT model. The GPCM is an IRT model suitable for polytomous data that allows the number of response options to differ between items

Table 1. Mapping of PF-10 and HAQ DI total scores on the latent IRT metric*

PF-10 score	Latent physical function score	HAQ DI score	Latent physical function score
0	−2.18	0.000	3.90
5	−1.45	0.125	2.59
10	−1.80	0.250	2.13
15	−1.68	0.375	1.71
20	−1.26	0.500	0.75
25	−1.08	0.625	0.69
30	−0.88	0.750	0.65
35	−1.34	0.875	−0.31
40	−0.92	1.000	−0.37
45	−0.39	1.125	−0.36
50	−0.02	1.250	−1.18
55	0.28	1.375	−1.21
60	0.04	1.500	−1.27
65	0.59	1.625	−1.68
70	0.84	1.750	−2.43
75	1.13	1.875	−2.82
80	1.72	2.000	−3.20
85	2.33	2.125	−3.58
90	2.75	2.250	−3.97
95	3.81	2.375	−4.35
100	4.59	2.500	−4.73
		2.625	−5.12
		2.750	−5.50
		2.875	−5.88
		3.000	−6.27

* Health Assessment Questionnaire disability index (HAQ DI) scores of >1.875 were imputed by linear regression, due to a paucity of data. PF-10 = Short Form 36 physical functioning scale; IRT = item response theory.

(15). This feature allowed the items of the different instruments to be concurrently calibrated despite their differing number of response options. The resulting IRT metric allows physical function levels for individual instruments to be estimated on a common scale so that their relative accuracy in recovering simulated physical function levels can be straightforwardly compared. The latent scale has a mean of 0, which is centered around the average level of physical function observed in the sample of RA patients that was used to calibrate the scale, and an SD of 1. Additional details regarding the calibration sample and model fit are described elsewhere (12).

CAT specifications. The CAT algorithm adaptively selects items from the PROMIS physical function item bank, and the specifications used in the current study correspond to the first-generation PROMIS CAT engine (16). In this study, Bayesian procedures were used; that is, the expected a posteriori method was used to obtain (interim) physical function estimates, and the maximum posterior-weighted information item selection criterion was used (17). The CAT algorithm was specified to assume a physical function level of $\theta = 0$ before the administration of items was started. Therefore, all simulated runs started with the item "Are you able to reach into a low cupboard?" (PFC31), which is the item with the highest precision at this level of physical function.

Data simulation and statistical analysis. The first set of simulations pertained to the relative measurement precision of the HAQ DI and PF-10 compared with a PROMIS physical function CAT algorithm with 5, 10, or 15 items. The ability of the different instruments to recover simulated "true" physical function scores was evaluated at 9 different levels of physical function ranging from $\theta = -4$ (i.e., 4 SD below the mean level of physical function in the calibration sample) to $\theta = 4$. For each of these score levels, 5,000 latent physical function scores (θ) were sampled from a normal distribution with an SD of 1. WinGen software (18) was used to generate response patterns consistent with the item parameters previously obtained by our group (12) and the specified levels of θ. Instrument-specific simulations were subsequently run in FireStar (19), and the difference between the actual specified level of θ and the

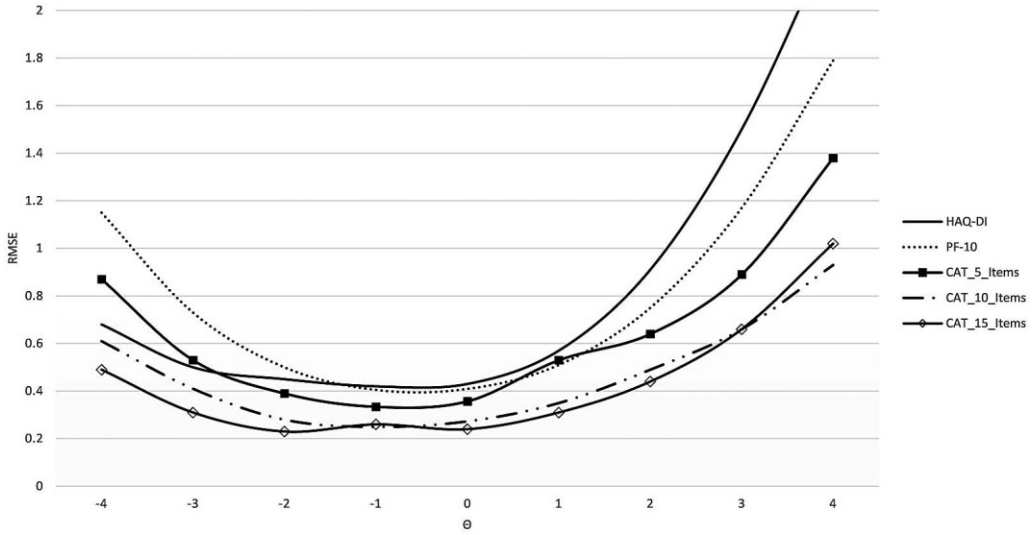

Figure 1. Conditional precision of physical function measures across various levels of the latent physical function scale. RMSE = root mean square error; HAQ DI = Health Assessment Questionnaire disability index; PF-10 = Short Form 36 physical functioning scale; CAT = Computer Adaptive Test; θ = latent physical function scale.

Table 2. Exposure rates of PROMIS physical function bank items included in the analysis of power*

Item code	Item	T1	T2
PFA8	Are you able to stand up from an armless straight chair?	192	267
PFA55	Are you able to walk a block (about 100 m) on flat ground?	671	330
PFA21	Are you able to step up and down curbs?	2,072	1,346
PFA42	Are you able to carry a laundry basket up a flight of stairs?	2	0
PFB9	Are you able to stand without losing your balance for several minutes?	34	14
PFA53	Are you able to dress yourself, including tying shoelaces and buttoning your clothes?	71	33
PFB10	Are you able to climb up five steps?	2,406	1,971
PFA6	Are you able to put on and take off a coat or jacket?	0	2,905
PFB14	Are you able to remove something from your back pocket?	869	605
PFC13	Are you able to get in and out of a car?	0	
PFA30	Are you able to pour liquid from a bottle into a glass?	1,080	566
PFB54	Does your health now limit you in going for a short walk (less than 15 minutes)?	49	26
PFB1	Are you able to reach into a low cupboard?	5,000	5,000
PFA36	Are you able to use a hammer to pound a nail?	1,012	526
PFB24	Are you able to lift one pound (0.5 kg) to shoulder level without bending your elbow?	827	422
PFB23	Are you able to stand for short periods of time?	392	199
PFA18	Are you able to go up and down stairs at a normal pace?	3,818	4,662
PFB33	Are you able to run errands and shop?	2,233	2,951
PFB5	Are you able to carry a shopping bag or briefcase?	0	19
PFC10	Are you able to stand up on tiptoes?	460	269
PFB56	Are you able to do yard work like raking leaves, weeding, or pushing a lawn mower?	0	534
PFC6	Are you able to walk at a normal speed?	46	25
PFA5	Are you able to move a chair from one room to another?	4,541	4,664
PFB44	Are you able to carry a laundry basket up a flight of stairs?	4,221	1,990
PFB48	Does your health now limit you in going OUTSIDE the home, for example to shop or visit a doctor's office?	1,871	1,453
PFB40	Are you able to wash and dry your body?	4,525	4,365
PFA37	Does your health now limit you in bathing or dressing yourself?	678	3,328
PFA25	Are you able to do chores such as vacuuming or yard work?	98	401
PFA11	Are you able to take a tub bath?	2,134	1,682
PFA15	Are you able to jump up and down?	2,784	2,813
PFA4	Does your health now limit you in lifting or carrying groceries?	240	722
PFA56	Are you able to run a short distance, such as to catch a bus?	863	1,309
PFA16	Does your health now limit you in climbing several flights of stairs?	391	887
PFB49	Does your health now limit you in doing moderate work around the house like vacuuming, sweeping floors or carrying in groceries?	1,028	1,945
PFB50	Does your health now limit you in doing moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf?	169	607
PFC12	Does your health now limit you in bending, kneeling, or stooping?	6	29
PFC39	Are you able to run 100 yards (100 m)?	1,368	2,062

* PROMIS = Patient-Reported Outcomes Measurement Information System; T1 = exposure rate at the baseline assessment; T2 = exposure rate at the second assessment.

estimates obtained from these runs was quantified by calculating the root mean square error (RMSE) for each instrument at each of the 9 levels of the latent physical function scale.

A second set of simulations was performed to evaluate the statistical power of the HAQ DI, PF-10, and PROMIS physical function 10-item CAT (CAT-10) to detect treatment effects on physical function that are typically observed in clinical trials. To allow assessment of the performance of the various instruments at levels that are likely to be seen in clinical trial settings, we reviewed a recent meta-analysis of the treatment effects of biologic agents on physical function in RA (20). The mean HAQ DI baseline score across the 28 studies included in the meta-analysis was close to 1.50, which corresponds to $\theta = -1.27$. Therefore, the means were set equal to $\theta = -1.27$ for the baseline analysis (T1) and to effect sizes of 0.20, 0.35, and 0.50, respectively, for the second assessment (T2). The means of the distribution of θ at the 2 time points were estimated using marginal maximum likelihood for a repeated-measures design (21). For this procedure, the autocorrelation across time points was set to equal -0.70. The second series of simulations was run with analogous settings as outlined above. All estimates and their SDs were obtained using Multidimensional Item Response Theory software (22). The ability of the instruments to detect the occurred changes in true θ was quantified by calculating the proportion of times the difference between the means at T1 and T2 was significant at a 1-sided alpha value of 5% for sample sizes of 25, 50, 75, and 100 patients. The statistical test used was the estimated

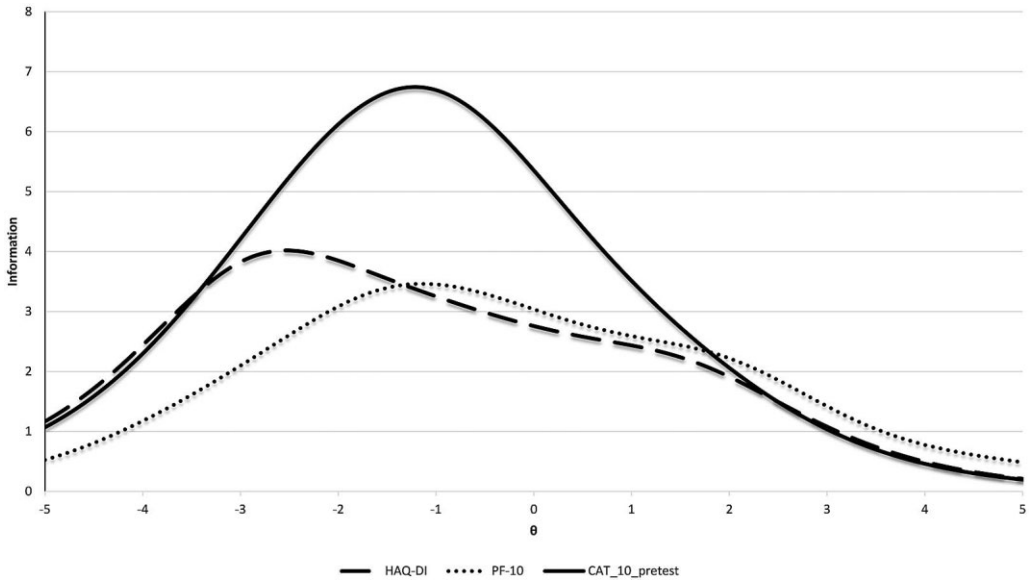

Figure 2. Local reliability of the 10 most frequently administered Patient-Reported Outcomes Measurement Information System (PROMIS) physical function items at $\theta = -1.27$ compared with the Health Assessment Questionnaire disability index (HAQ DI) and the Short Form 36 physical functioning scale (PF-10). Information = measure of local reliability defined as the inverse square of the standard error of measurement for each level of θ (SE $\theta = 1/\sqrt{([I(\theta)])}$); θ = latent physical function scale; CAT_10_pretest = 10 most frequently selected PROMIS physical function items by CAT algorithm at $\theta = -1.27$.

difference between the means divided by the SD of this difference. This statistic was considered to follow a normal distribution.

RESULTS

Table 1 shows mapping of physical function scores for the HAQ DI and PF-10, obtained by applying the standard scoring algorithms on the latent IRT-based metric. Despite the different manners in which the scores are obtained and expressed, IRT- and summed score–based approaches generally yield highly correlated scores. The correlations between IRT-based and standard scores were 0.97 and 0.95 for the HAQ DI and PF-10, respectively. Hereafter, results will be reported for the IRT metric only, but Table 1 can be used as a frame of reference for those more familiar with the standard scoring algorithms for these questionnaires.

Measurement precision. Figure 1 shows the conditional precision of the instruments across 9 levels of physical function. The most precise estimates of physical function were achieved at physical function levels at the average of the scale of 1 SD below the average. In general, the precision of all instruments was better for lower levels of function, and measurement performance decreased sharply for physical function levels ≥2 SD above average, with, for example, RMSE = 0.44 for the CAT-15. The performance of the HAQ DI was optimal for lower levels of physical function, and its precision remained relatively constant for below-average physical function levels. In contrast, the PF-10 performed com-

Table 3. Percentage of 1,000 simulated clinical trials in which the null hypothesis of equal means across time points was rejected at Z = 1.645*

Sample size, no.	Effect size 0.20		PROMIS CAT-10	Effect size 0.35		PROMIS CAT-10	Effect size 0.50		PROMIS CAT-10
	PF-10	HAQ DI		PF-10	HAQ DI		PF-10	HAQ DI	
25	5.8	8.4	4.3	15.5	17.5	19.5	51.4	39.5	61.8
50	12.9	11.9	11.2	36.1	40.2	49.8	85.1	75.6	93.9
75	19.9	18.1	18.3	55.1	58.9	72.5	95.7	94.2	99.8
100	28.3	27.8	26.1	68.1	73.9	86.5	99.5	98.8	100

* PF-10 = Short Form 36 physical functioning scale; HAQ DI = Health Assessment Questionnaire disability index; CAT-10 = Computer Adaptive Testing algorithm using 10 items.

paratively poorly at the lowest levels of physical function but outperformed the HAQ DI at above-average levels of physical function. Interestingly, a CAT with as few as 5 items outperformed both the HAQ DI and PF-10 across most levels of function. The CAT-10 and CAT-15 algorithms produced the most accurate and very similar results across all levels of physical function, with the CAT-15 having marginally higher precision only for very low levels of physical function.

Detection of treatment effects. In the analysis of power for an effect size of 0.50, 45 of 121 PROMIS physical function items were chosen by the CAT-10. The item exposure rates, ranked from least to most difficult to perform, are shown in Table 2. Generally, more difficult items had higher exposure rates at T2, consistent with the improved level of simulated physical function at that time point. The test information function of the 10 items with the highest exposure rates at T1 are plotted in Figure 2, together with the test information functions of the PF-10 and the HAQ DI. Test information functions describe the local reliability of the measures across the latent continuum of physical function. The HAQ DI had higher overall reliability than the PF-10 for lower levels of physical function, which corresponds to the results of the previous analyses. Furthermore, the CAT-10 was successful at optimizing measurement precision at $\theta = -1.27$. The measurement precision of the 10 most frequently chosen items was in fact higher than that of both classic instruments across most levels of physical function. This likely reflects the extra information provided by the greater number of response options of the PROMIS physical function items compared with the HAQ DI or PF-10 items.

Table 3 shows the results of analysis of the power of the instruments to detect treatment effects of various magnitudes and sample sizes. The performance of the 3 measures for detecting small treatment effects was very similar, with the exception of a higher detection rate for the HAQ DI, at a sample size of 25 patients. For effect sizes of 0.35 and 0.50, the PROMIS physical function CAT-10 clearly outperformed both classic measures. This likely reflects that notion that the higher latent physical function scores at T2 enable the CAT algorithm to select items from the range in which the item bank is richest in information (i.e., the range of latent physical function scores between -1 and 0).

DISCUSSION

Applications of IRT-based item banking such as CATs are potentially useful tools that facilitate the assessment of the broad range of physical function levels observed in RA, without the need to resort to extensively long fixed-length questionnaires. The current study assessed the measurement precision and efficiency of the commonly used fixed-length HAQ DI and PF-10 compared with a CAT algorithm based on the PROMIS physical function item bank, using simulation studies. Overall, the findings suggest that a 10-item PROMIS physical function CAT performs better than the HAQ DI and PF-10 in terms of measurement precision and sensitivity to change.

The results of the current study illustrate that the items of the HAQ DI are slightly better suited for relatively poorly functioning patients, and that the PF-10 items are slightly better suited for well-functioning populations. These results correspond to previous work in RA, in which substantial ceiling and floor effects, respectively, were observed for these measures (5,6). In fact, it has also been demonstrated using IRT-based methodology that more favorable scaling properties with respect to ceiling and floor effects could be achieved in RA if modified HAQ and PF-10 items were concurrently calibrated (23). Administering multiple fixed-length measures of physical function to extend the range of physical function levels, however, is not a practical solution to the scaling problems associated with the assessment of physical function in RA, considering the added burden this would entail for patients and persons who wish to obtain physical function data for research or clinical purposes.

However, before the use of CATs in clinical settings can be recommended, it is necessary to evaluate the performance of CAT applications in comparison with the performance of existing measures that are routinely used in current clinical studies and that have proven their worth over time (24). It has previously been shown that a PROMIS physical function short form was highly reliable, and that smaller sample sizes would be needed for clinical studies using this measure (25). The current study extends these findings to CATs using the PROMIS physical function item bank and demonstrates that large improvements in power could be obtained by using CATs with a small number of items. Specifically, 5-item CATs yielded superior precision compared with both of the fixed-length questionnaires in recovering simulated physical function scores across all but the very lowest levels of physical function, and 10-item and 15-item CATS yielded superior precision across all evaluated levels of physical function. Similar results were obtained with a preliminary version of the PROMIS

physical function item bank based on a calibration of historical physical function data (26).

These findings suggest that impressive gains in the efficiency of physical function assessment can be achieved in practice by using CAT assessments based on the PROMIS physical function item bank, without sacrificing measurement precision. At below-average levels of physical function, the HAQ DI performed reasonably well, reflecting its items measuring mostly simple activities of daily living, while the largest differences in measurement precision between the PROMIS physical function CAT and both classic instruments were observed at above-average levels of physical function. Relatively well-functioning patients with RA are an increasingly relevant population in light of the current overall improved status of these patients (27). For example, there has recently been an increase in studies evaluating the possibility of down-titrating anti–tumor necrosis factor treatment or disease-modifying antirheumatic drugs in patients in whom remission is stable (28,29).

Clinical studies in RA typically show improvements in physical function of moderate to large magnitude (i.e., effect size $\geq$0.50) (20). In the present study, almost 95% of the simulated moderate effects were correctly identified by statistical tests based on the estimates obtained with the PROMIS physical function CAT-10, at a sample size of 50 patients. To achieve similar power with the HAQ DI or PF-10, 75 patients would have to be included in the experimental arm. These results have important practical implications, because the inclusion of fewer patients in clinical trials saves costs. An important point to note here is that the number of patients needed to achieve a certain power was higher than would be suggested by power tables that are routinely used for determining sample size in clinical studies. The main reason for this phenomenon is that when using a *t*-test or similar statistical technique, it is assumed that the end point is measured without error. However, the estimation error associated with the IRT estimates must be taken into account. Therefore, the reliability of the instrument is an extra variable that should be considered when determining the sample size for a study. This is accomplished by using the marginal maximum likelihood method to estimate the means of the latent variable distribution at the 2 time points (21).

Although the results of the current study are highly encouraging, limited availability of technology and limited familiarity with CATs are potential bottlenecks in the adoption of CATs in current research practice (30). Researchers interested in using the PROMIS physical function item bank in RA in settings where computerized assessment is not feasible may choose to administer a subset of the items that were most frequently chosen in the evaluation of power (see Table 3). These items are specifically tailored to the average level of physical function experienced by patients with RA at the time of entry into clinical trials, and the 10 most frequently chosen items were shown to represent a more reliable assessment of physical function than the 20 items of the HAQ DI or the PF-10, not only with respect to the relevant levels of physical function but also across all but the very highest levels of physical function. The obtained item response data can then be scored using the item parameters available at the PROMIS assessment center (www.assessmentcenter.net), so that results can be expressed on the standardized PROMIS metric. This allows the impact of RA on physical function to be compared with that of other conditions (31).

A general limitation of the use of short forms instead of CATs is that they are not optimally tailored to individual patients, resulting in lower-than-optimal reliability for patients with unusual levels of function and possible floor and ceiling effects. However, manually selecting fixed items to be administered is one way to ensure the content validity of the assessment. In the CAT algorithms used in the current study, items were selected based purely on their measurement precision without considering the content of the individual items. In contrast, the HAQ DI has 8 categories of physical function that are all of particular relevance to RA, which strengthens its content validity for RA. Content constraints may also be imposed on the CAT selection algorithm, which will ensure a balanced assessment of various aspects of physical function that are deemed relevant a priori (32).

The item parameters used in the present study were obtained from a sample of Dutch patients with RA, whereas the original PROMIS physical function item parameters were calibrated in a sample of the general population in the US. In a previous study, we demonstrated that although some (20%) of the PROMIS physical function items functioned differently when comparing Dutch RA patients with the US general population, this did not cause bias in total physical function estimates (12). Similar results were obtained in a study of the equivalence of Dutch and US physical function data in RA, in which the presence of 2 of 10 items with differential item functioning (DIF) had a negligible impact on the total scores (33). Considering that the majority (>80%) of items that were adminis-

tered in the CAT simulations in the current study were shown to be free of DIF, it is unlikely that different results would have been obtained using the US PROMIS calibration for CAT-10 and CAT-15. However, no previous studies are available that assessed the impact of DIF on total scores with applications using only 5 items. Therefore, the results of CAT-5 should be interpreted with caution, and it is important that future studies further evaluate the validity of the PROMIS parameters for use in specific applications in RA.

An inevitable limitation of the simulation methodology used in the current study is the assumption that the model holds perfectly. The validity of a post hoc simulation study can in principle be confirmed by performing empirical simulation studies that use real data to simulate the CAT responses. Due to the research design that was used to obtain the parameters for the PROMIS physical function items in both the US and Dutch calibration studies, patients responded to only a limited number of items, precluding a realistic empirical simulation. However, previous studies with polytomous data have generally detected small differences between empirical and post hoc simulations (33).

In summary, the current study demonstrates the considerable gains in efficiency and measurement precision that may be achieved in the assessment of physical function in RA using CATs based on the PROMIS physical function item bank. Overall, a 10-item CAT demonstrated better measurement precision compared with classic fixed-length questionnaires, which may translate to increased power to detect treatment effects in clinical studies in RA.

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 published. Dr. Oude Voshaar 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. Oude Voshaar, ten Klooster, Glas, Vonkeman, Krishnan, van de Laar.

Acquisition of data. Oude Voshaar, ten Klooster, Vonkeman, van de Laar.

Analysis and interpretation of data. Oude Voshaar, ten Klooster, Glas, Vonkeman.

REFERENCES

1. Wolfe F, Lassere M, van der Heijde D, Stucki G, Suarez-Almazor M, Pincus T, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:484–9.
2. Van der Heijde D. Impact of rheumatoid arthritis on physical function during the first five years. No longer a question mark? Rheumatology (Oxford) 2000;39:579–80.
3. Kalyoncu U, Dougados M, Daures JP, Gossec L. Reporting of patient-reported outcomes in recent trials in rheumatoid arthritis: a systematic literature review. Ann Rheum Dis 2009;68:183–90.
4. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
5. Oude Voshaar MA, ten Klooster PM, Taal E, van de Laar MA. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes 2011;9:99.
6. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis 1995;54:461–5.
7. Ni P, McDonough CM, Jette AM, Bogusz K, Marfeo EE, Rasch EK, et al. Development of a computer-adaptive physical function instrument for Social Security Administration disability determination. Arch Phys Med Rehabil 2013;94:1661–9.
8. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al, PROMIS Cooperative Group. 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.
9. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol 2006;59:1174–82.
10. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park (CA): Sage Publications; 1991.
11. Van der Linden WJ, Glas CA. Computerized adaptive testing: theory and practice. Dordrecht (The Netherlands): Kluwer Academic Publishers; 2000.
12. Oude Voshaar MA, ten Klooster PM, Glas CA, Vonkeman HE, Taal E, Krishnan E, et al. Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. PLoS One 2014;9:e92367.
13. Ware JE Jr, Sherbourne CD. The MOS 36-Item Short-Form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
14. Terwee CB, Roorda LD, de Vet HC, Dekker J, Westhovens R, van Leeuwen J, et al. Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). Qual Life Res 2014;23:1733–41.
15. Muraki E. A generalized partial credit model: application of an EM algorithm. Appl Psychol Meas 1992;16:159–76.
16. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res 2010;19:125–36.
17. Van der Linden W. Bayesian item selection criteria for adaptive testing. Psychometrika 1998;63:201–16.
18. Han KT. WinGen: Windows software that generates item response theory parameters and item responses. Appl Psychol Meas 2007;31:457–9.
19. Choi SW. Firestar: computerized adaptive testing simulation program for polytomous item response theory models. Appl Psychol Meas 2009;33:644–5.
20. Barra L, Ha A, Sun L, Fonseca C, Pope J. Efficacy of biologic agents in improving the Health Assessment Questionnaire (HAQ) score in established and early rheumatoid arthritis: a meta-analysis with indirect comparisons. Clin Exp Rheumatol 2014;32:333–41.
21. Glas CA. Preliminary manual of the software program Multidimensional Item Response Theory (MIRT). 2010. URL: http://www.utwente.nl/gw/omd/Medewerkers/temp_test/mirt-manual.pdf.
22. Martin M, Kosinski M, Bjorner JB, Ware JE Jr, MacLean R, Li T.

Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647–60.
23. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
24. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61:17–33.
25. Welsing PM, Fransen J, van Riel PL. Is the disease course of rheumatoid arthritis becoming milder? Time trends since 1985 in an inception cohort of early rheumatoid arthritis. Arthritis Rheum 2005;52:2616–24.
26. Van Herwaarden N, Den Broeder A, Jacobs W, Bijlsma JW, Van Vollenhoven RF, Van den Bemt BJ. Down titration and discontinuation strategies of tumor necrosis factor blocking agents for rheumatoid arthritis in patients with low disease activity (protocol). The Cochrane Library, 2013. URL: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD010455/pdf.
27. O'Mahony R, Richards A, Deighton C, Scott D. Withdrawal of disease-modifying antirheumatic drugs in patients with rheumatoid arthritis: a systematic review and meta-analysis. Ann Rheum Dis 2010;69:1823–6.
28. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther 2011;13:R147.
29. Glas CA, Geerlings H, van de Laar MA, Taal E. Analysis of longitudinal randomized clinical trials using item response models. Contemp Clin Trials 2009;30:158–70.
30. Revicki DA, Sloan J. Practical and philosophical issues surrounding a national item bank: if we build it will they come? Qual Life Res 2007;16 Suppl 1:167–74.
31. Rothrock NE, Hays RD, Spritzer K, Yount SE, Riley W, Cella D. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2010;63:1195–204.
32. Veldkamp BP, van der Linden WJ. Multidimensional adaptive testing with constraints on test content. Psychometrika 2002;67:575–88.
33. Makransky G, Mortensen EL, Glas CA. Improving personality facet scores with multidimensional computer adaptive testing: an illustration with the NEO PI-R. Assessment 2013;20:3–13.