ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 1175−1183

Journal of Clinical Epidemiology

Items and dimensions for the construction of a multidimensional computerized adaptive test to measure fatigue in patients with rheumatoid arthritis

Stephanie Nikolaus[a,*], Christina Bode[a], Erik Taal[a], Johanna C.M. Oostveen[b], Cees A.W. Glas[c], Mart A.F.J. van de Laar[a,d]

[a]*Department of Psychology, Health & Technology, University of Twente, Faculty of Behavioral Sciences, P.O. Box 217, 7500 AE Enschede, The Netherlands*
[b]*Department of Rheumatology, ZGT Ziekenhuisgroep Twente, Almelo, The Netherlands*
[c]*Department of Research Methodology, Measurement and Data Analysis, University of Twente, Faculty of Behavioral Sciences, Enschede, The Netherlands*
[d]*Department of Rheumatology and Clinical Immunology, Medical Spectrum Twente, Enschede, The Netherlands*

Accepted 2 May 2013; Published online 17 August 2013

Abstract

Objectives: Development of an item pool to construct a future computerized adaptive test (CAT) for fatigue in rheumatoid arthritis (RA). The item pool was based on the patients' perspective and examined for face and content validity previously. This study assessed the fit of the items with seven predefined dimensions and examined the item pool's dimensionality structure in statistical terms.

Study Design and Setting: A total of 551 patients with RA participated in this study. Several steps were conducted to come from an explorative item pool to a psychometrically sound item bank. The item response theory (IRT) analysis using the generalized partial credit model was conducted for each of the seven predefined dimensions. Poorly fitting items were removed. Finally, the best possible multidimensional IRT (MIRT) model for the data was identified.

Results: In IRT analysis, 49 items showed insufficient item characteristics. Items with a discriminative ability below 0.60 and/or model misfit effect sizes greater than 0.10 were removed. Factor analysis on the 196 remaining items revealed three dimensions, namely severity, impact, and variability of fatigue. The dimensions were further confirmed in MIRT model analysis.

Conclusion: This study provided an initially calibrated item bank and showed which dimensions and items can be used for the development of a multidimensional CAT for fatigue in RA.

Keywords: Fatigue; Rheumatoid arthritis; Item response theory; Patient-reported outcome; Psychometrics; Computerized adaptive test

1. Introduction

Many patients with rheumatoid arthritis (RA) complain about fatigue [1,2]. However, the causes of fatigue in patients with RA are not yet fully understood [3]. Patients experience fatigue as a multidimensional, annoying symptom with far-reaching consequences [4−7] and report to receive no adequate medical or paramedical support for it [8]. They describe their fatigue being different from normal tiredness, as it is often more extreme, not always caused by high levels of activity and therefore unpredictable [5].

Measuring fatigue provides important information for understanding the patients' perspective on disease impact and outcome [9]. The measurement with single item scales has some but limited value [9] because it does not correspond to the multidimensional character of fatigue [4−7]. Instead, there are several established multi-item fatigue questionnaires, for example the Chalder fatigue scale [10], Multidimensional Fatigue Inventory [11], Short Form-36 subscale vitality [12], Functional Assessment Chronic Illness Therapy (Fatigue) [13], Profile of Mood States subscale fatigue/inertia [14], or the Checklist Individual Strength [15]. However, it is disputable how appropriate these instruments are for measuring fatigue in RA because none of them was specifically developed for an

This study was financed by the Foundation of Rheumatism Research Twente (Stichting Reumaonderzoek Twente) and the IBR Research Institute for Social Sciences and Technology and conducted at the Arthritis Centre Twente; a collaboration between the University of Twente and Medical Spectrum Twente.

Financial conflict of interest: None.

* Corresponding author. Tel.: +31-534896063; fax: +31-53-4892895.
E-mail address: s.nikolaus@utwente.nl (S. Nikolaus).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.05.010

What is new?

Key findings

- The multidimensional character of fatigue as reported by patients with rheumatoid arthritis (RA) is also supported by item response theory (IRT); a three-dimensional IRT model provided the best possible fit to the data.

What this adds to what was known?

- The uniqueness of this study lies in the combination of the patients' perspective and modern psychometrics in the development of a multidimensional item bank to measure fatigue in RA.

What is the implication and what should change now?

- Our initially calibrated item bank will be used for the development of a multidimensional computerized adaptive test, aiming to obtain a precise and efficient instrument to measure fatigue in RA that can be used in different settings.

RA population. As a result, generic fatigue items might be confounded by disease-specific conditions [16,17]. In addition, not all multi-item fatigue questionnaires measure fatigue as a multidimensional construct. Although one established multidimensional RA-specific instrument does exist, the Multidimensional Assessment of Fatigue [18], its content validity is questionable as it was not developed from the patients' perspective. The experience of patients is essential in the development of questionnaire items to ensure content validity [19,20] because only they can report on the subjective experience of fatigue [21]. Especially for measuring such a complex phenomenon as fatigue, where underlying causal pathways are not yet fully understood, the view of patients should form the basis. Furthermore, as long as it is not clear whether or how fatigue in RA differs from fatigue in other medical conditions, it is useful to start with disease-specific measures [16]. The Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire [22] is a relatively new questionnaire that meets these requirements. It is an RA-specific questionnaire that includes the patients' perspective and is multidimensional, and is currently undergoing extensive validation.

However, common to all the previously mentioned fatigue scales is that they have a traditional, fixed length format. This is time consuming because patients have to answer questions that may not apply to their situation. Furthermore, existing questionnaires do not capture all aspects of fatigue [23]. So more appropriate and efficient ways of measurement are needed.

Our ultimate aim is to develop a new multidimensional instrument for fatigue in RA that is based on the patients' perspective and also uses the advantages of modern measurement technology.

Computerized adaptive testing (CAT) provides the possibility to comprehensively measure patient-reported outcomes (PROs) with few items [24]. The CAT is an upcoming domain in medical settings [25]. The CATs for depression, anxiety, and stress perception turned out to be reliable, valid, and efficient instruments that measure more precisely than traditional questionnaires [26−28]. Using CAT decreases the burden for patients because not everybody has to answer all the same questionnaire items. It also increases measurement precision as items are sequentially selected from an item bank based on the previous answer of this patient.

However, for the computerized selection of the best matching items, a calibrated item bank that contains much more items than are finally presented to a single patient should be developed first [29], which was the aim of the present study. Before a CAT can be developed, an item bank has to be scaled with item response theory (IRT). Item parameters as the difficulty level can be estimated for each individual item and the scale values for fatigue levels [29]. Consequently, we can estimate the level of fatigue reflected by the item and all items are placed on this continuum, ranging from no fatigue to severe fatigue. Furthermore, it can be calculated how well an item discriminates between more or less fatigued patients. This information is required to optimally match the items to the patient's individual level and support interindividual comparisons on the measured construct even if patients filled in different items. Because of the multidimensional nature of fatigue, a between-items multidimensional IRT (MIRT) model will be used. In a between-items MIRT model, it is assumed that the item bank pertains to a limited number of correlated latent dimensions (say three) and that every item loads on one dimension only. Through the correlation between the latent dimensions, item responses provide information regarding the position of a patient in the latent space. This approach provides more information than the approach measuring each dimension separately.

In preparation of the development of the calibrated item bank, we constructed an item pool based on the patients' perspective. To capture all relevant aspects of fatigue, we first investigated the experience of fatigue [7,30]. Then, we collected items and dimensions of existing fatigue scales and supplemented them with items from interview material [7] in a preliminary item pool. This item pool was evaluated in a Delphi study with Dutch experts (patients, nurses, and rheumatologists) to select adequate items to measure fatigue in RA [31−33]. The final content valid item pool consisted of 245 items and 12 dimensions as shown in the first column of Fig. 1. This flowchart shows the item selection process throughout this study.

The difference between our research and already existing IRT approaches in the field of fatigue is that we do

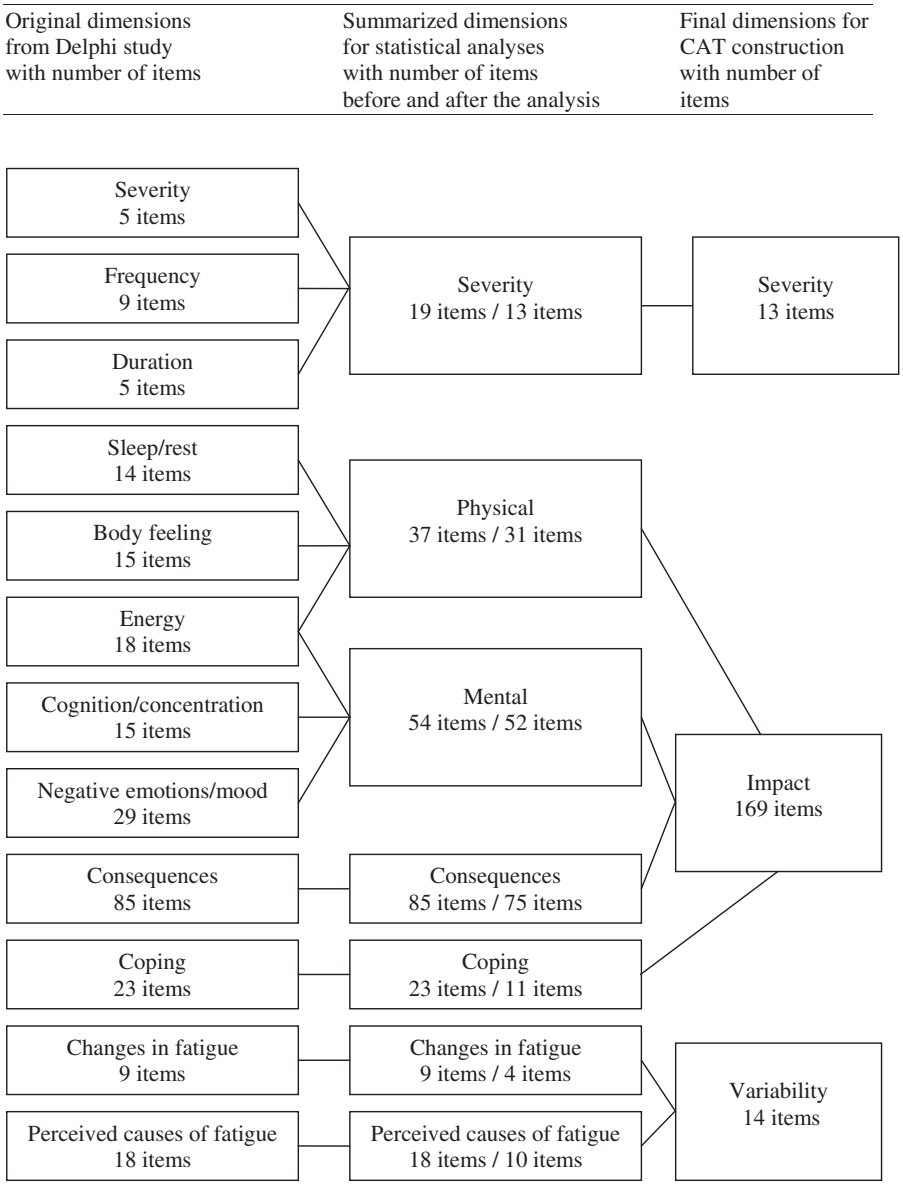

Fig. 1. Flowchart of the item selection process. CAT, computer adaptive test.

not intend to develop one item bank for each dimension of fatigue, but to produce a multidimensional instrument. A joint initiative is currently developing a Patient-Reported Outcome Measurement Information System that aims to construct a large item bank and CAT systems for assessing PROs in chronic diseases [34]. Within that scope, two separately calibrated fatigue item banks (experience and impact) were developed for use in the general population and different chronic conditions [35]. Our aim, in contrast, was to calibrate one large item bank that contains several dimensions together and that is developed based on the perspective of patients with RA.

2. Patients and methods

2.1. Patients

Altogether, 999 patients with RA of the Ziekenhuis Groep Twente (Almelo and Hengelo) and the Arthritis Center Twente at Medical Spectrum Twente (Enschede) were invited for participation. They received a letter from their rheumatologist, informing about the study aim and details about participation. Each letter was accompanied by one version of the fatigue questionnaire, an informed consent form, and a paid return envelope. The data collection took place in 2011.

2.2. Questionnaire

2.2.1. Patient characteristics

The questionnaire started with demographic characteristics (gender, date of birth, marital status, level of education, and work status) and disease characteristics (year of disease onset and comorbidities).

2.2.2. Disease-related measures

Patients filled in 11-point numerical rating scales (NRSs) for pain and impact of the disease, and three fatigue NRS [22] for severity, impact, and coping. As these three NRSs were developed in the United Kingdom, they were translated into Dutch [36]. Furthermore, patients completed the 10-item version of the Health Assessment Questionnaire-II [37].

2.2.3. Fatigue items

Owing to the size of our item pool, it was not feasible to let each participant fill in all fatigue items. The burden of answering 245 questions would be too high, so we prepared different versions of questionnaires containing between 103 and 126 items. This resulted in a common-item linking design [38]. Each questionnaire version not only consisted of a different composition of dimensions and items but also had some sets of items in common (see Table A1 in Appendix A). The common-item linking design was developed in such a way that the items and dimensions of the different questionnaire versions could be related to each other in the IRT analysis [38]. All the parameters in the IRT model were concurrently estimated using a marginal maximum likelihood procedure for data collected in a common-item linking design [39].

Our previous Delphi study revealed 12 content valid dimensions of fatigue [32] as displayed in the first column of Fig. 1. However, for the construction of the linking design, these were too many. Some of the dimensions are closely related to each other (e.g., severity, frequency, and duration are all about the manifestation of fatigue) and are measured as one dimension in other fatigue questionnaires. They were combined in the further analyses (see the second column of Fig. 1). In contrast, other dimensions are not yet frequently covered by fatigue scales (changes in fatigue, perceived causes of fatigue, and coping with fatigue) and therefore separately included in the further analysis and included in most of the questionnaire versions. By this means, we wanted to find out how these dimensions, consisting of many newly constructed items, would fit in the IRT model.

2.3. Analyses

The ultimate aim is to build a CAT based on a between-items multidimensional IRT model. The straightforward search for such a model with MIRT tools is complicated. Therefore, preliminary analyses using factor analysis are frequently conducted. A traditional factor analysis based on a tetrachoric correlation matrix is comparable with an IRT analysis under certain circumstances, but this is not generally the case [40]. Nevertheless, these analyses produce a good initial indication of the dimensionality structure for the MIRT model. So, three preparatory steps were made to obtain information about the dimensionality structure for fitting the final MIRT model in a forth step.

2.3.1. Step 1: initial item selection

For each of the seven larger dimensions (Fig. 1), the fit to a unidimensional IRT model and the measurement quality were investigated. For each of the seven dimensions, an IRT analysis was conducted with public domain software MIRT [39] under the generalized partial credit model (GPCM) [41]. This model is applicable to analyze polytomous items, meaning items with more than two response options, and allows the items within a scale to differ in discrimination parameter values [42]. The discrimination parameter is highly correlated with the item/rest score [40]. The item/rest score correlation is the correlation between a specific item response and the total score without the specific item. In classical test theory, it is used as an indication for the contribution of the item to the reliability of the test. We dismissed all items with a discrimination parameter below 0.60. Such items contribute little to the overall reliability and will not be selected in the CAT anyway. A low discrimination parameter means that an item does not discriminate well between more or less fatigued persons. Furthermore, model fit to the IRT model was investigated using Lagrange multiplier (LM) test statistics [43]. We excluded items where the effect size of the misfit, that is the size of the difference between the average observed and expected item score in subgroups, was larger than 0.10 [44,45].

2.3.2. Step 2: IRT analysis of the fit for the separate predefined dimensions after removal of the unreliable items

We repeated the IRT analyses carried out in step 1, but without the excluded items. Again, the LM test statistic outlined in step 1 was used. The number of significant LM model tests and reliabilities as estimated under the GPCM, are provided in the Results section. If the data strictly fit the model, the percentage of model tests that are significant at the 5% level should also be approximately 5% of all the model tests.

2.3.3. Step 3: initial exploration of the dimensionality structure of the item pool

The data set has too many items ($K = 196$) to run exploratory analyses in standard software such as Mplus (Muthen & Muthen, Version 5.21). Therefore, further analyses were conducted on IRT-based person parameter estimates. To account for measurement error, the seven dimensions were split up into two parts, containing odd and even items, and plausible values of the person parameters were drawn for both sets [46]. So the input for MPlus consisted of 14 variables for each patient. For the interpretation of model

fit, the root mean square error of approximation (RMSEA) was used as criterion. An RMSEA value smaller or equal to 0.05 indicates a close fit between the observed correlation matrix and the correlation matrix expected under the model. Values between 0.05 and 0.08 suggest a reasonable approximation.

2.3.4. Step 4: confirmative MIRT analysis using the results of step 3

In step 4, the results of the exploratory analysis in step 3 were used to build an MIRT model that is suited for a first estimation to run the CAT. To evaluate the fit to the MIRT model, the same LM test statistics as used in the steps 1 and 2 were computed, and counts of significant LM tests were used.

3. Results

3.1. Participants

We received 551 completed questionnaires, a response rate of 55%. The sample consisted of 367 women, 164 men and 20 persons who did not fill in their sex. The level of fatigue, measured with the NRS, was 4.89 (standard deviation = 2.3), ranging from 0 to 10, showing that a broad range of fatigue was represented among the participants. Further sample characteristics are shown in Tables 1 and 2.

3.2. Development of the MIRT model

In this section, the results of the four analysis steps are described. The first three steps are intended to search for the structure of the MIRT model.

3.2.1. Step 1: initial item selection

We excluded 49 items owing to unsatisfactory item characteristics, that is, a discriminative ability below 0.60 and/or an misfit effect size larger than 0.10. Table A1 in Appendix A shows the number of excluded items per dimension. Furthermore, an overview table is provided as online supplementary material (Table S1 in Appendix B at www.jclinepi.com), showing the excluded items with abbreviated item content, factor loadings, and reasons for exclusion.

Table 1. Sample characteristics ($N = 551$)

Characteristics	Mean (SD)	Range
Age, yr	63.38 (12.70)	24–92
RA disease duration, yr	15.15 (11.22)	0–67
NRS items		
General health	4.54 (2.11)	0–10
Pain	4.38 (2.38)	0–10
Fatigue severity	4.89 (2.30)	0–10
Impact of fatigue	4.62 (2.53)	0–10
Coping with fatigue	6.50 (2.02)	0–10
HAQ-II score	1.00 (0.65)	0–3

Abbreviations: SD, standard deviation; RA, rheumatoid arthritis; NRS, numerical rating scale; HAQ, health assessment questionnaire.

Table 2. Sample characteristics ($N = 551$)

Characteristics	N
Sex	
Women	367
Men	164
Marital status	
Single	27
Living with partner/married	412
Widow/widower	73
Divorced	33
Level of education	
Low (≤12 yr of education)	362
Moderate (13–14 yr of education)	109
High (≥14 yr of education)	71
Work status	
Working full-time	58
Working part-time	89
Household/unemployed	105
Disabled/retired	290
Comorbidities	
Yes	260
No	291

3.2.2. Step 2: IRT analysis of the fit for the separate predefined dimensions after removal of the unreliable items

We counted the number of significant model tests among the different questionnaire versions per dimension as provided by Lagrange tests for GPCM. They are provided in Table 3. Also the reliabilities as estimated under the GPCM are shown per dimension.

The percentages of significant model tests are too high for the "severity" and "physical" dimensions, so for these two dimensions unidimensionality was not supported. The percentages of significance probabilities for the dimensions "mental," "consequences," and "change" were quite close to the nominal significance probability of 5%, so here unidimensionality was considered acceptable.

3.2.3. Step 3: initial exploration of the dimensionality structure of the item pool

Factor analysis was used as a tool that aids the search for the final IRT model. Four factor solutions were taken into account:

Table 3. Significant model test and reliability per dimension (estimated with GPCM)

Dimension	Significant model tests at 5%	Reliability
Severity	24/104 (23.1)	0.959
Physical	85/169 (50.3)	0.975
Mental	19/224 (8.5)	0.974
Consequences	19/307 (6.2)	0.978
Change	4/45 (8.9)	0.701
Perceived causes	13/79 (16.5)	0.646
Coping	12/88 (13.7)	0.878

Abbreviation: GPCM, generalized partial credit model.

1. A factor solution with one dimension was rejected. The analysis resulted in an RMSEA of 0.098. The maximal bound for the RMSEA is usually taken as 0.05. The test of the hypothesis that the RMSEA is smaller than 0.05 was highly significant: p(RMSEA $\leq 0.05) < 0.001$.
2. A model with two factors showed a good fit (RSMEA = 0.049, p[RMSEA ≤ 0.05] = 0.485), but the dimensions were hard to interpret, and the subsequent confirmatory MIRT analysis did not support between-items multidimensionality. All items loaded to some degree on both dimension, resulting in an uninterpretable within-items MIRT model.
3. Three factors fitted well and lead to RMSEA = 0.041 (p[RMSEA ≤ 0.05] = 0.534). This solution can also be interpreted in theoretical terms. The first factor consists of the predefined dimension 1 (severity of fatigue), the second factor consists of dimension 2 (physical), 3 (mental), 4 (consequences), and 7 (coping)—all dimensions referring to the impact/consequences of fatigue in a broader sense, and the third factor consists of dimensions 5 (changes) and 6 (perceived causes). These two latter dimensions contain several new formulated items and refer to aspects of the variability of fatigue.
4. Four dimensions (with the third dimension split into two dimensions: 5 [changes] and 6 [perceived causes]) did not result in better model fit. That is, the likelihood ratio test of a model with three dimensions against a model with four dimensions had a chi-square value of 2.133 with three degrees of freedom. That is, using four dimensions did not significantly improve model fit.

3.2.4. Step 4: confirmative MIRT analysis using the three-dimensional model

The initial three-factor solution of step 3 was used here to test the final MIRT model. The three-dimensional IRT model was compared with a one-dimensional GPCM using a likelihood ratio test. The value of the chi-square was 148, with two degrees of freedom, so the unidimensional model was clearly rejected. Analogous to the test of model fit of the predefined dimensions, fit to the IRT model was evaluated using counts of significant item tests. This resulted in 90 tests significant at 5% of the 597 tests conducted (15%). This outcome is not perfect but clearly a sufficient basis for the development of a CAT and for a first estimate to run this measurement instrument. The correlations of the latent variables in the multidimensional GPCM are shown in Table 4. The correlations are moderate.

3.3. Conclusion of the data analysis

The initially calibrated multidimensional item pool consists of 196 items, spread among three dimensions, namely severity (severity), impact (physical, mental, consequences, and coping), and variability (change and perceived causes), as displayed in the third column of Fig. 1. The online supplementary material includes a table (Table S2) showing for all 196 items the affiliated dimension, abbreviated item content, item source, minimum and maximum threshold, and slope parameters.

Table 4. Estimated correlations between the three dimensions

Dimensions	Severity (1)	Impact (2)	Variability of fatigue (3)
1	1.000	0.495	0.247
2		1.000	0.580
3			1.000

4. Discussion

This study provided the first calibrated item pool for the development of a multidimensional CAT for fatigue in patients with RA.

The strength of the item pool lies in its stepwise development from the patients' perspective and the thorough selection of meaningful items and dimensions. Before the statistical analyses described in this article, our item pool consisted of 245 items and 12 dimensions that were qualitatively evaluated by an expert panel [31−33].

These dimensions were already summarized into seven larger categories for the construction of the linking design that we used for the composition of the different questionnaire versions. In the first two steps, IRT analysis was used for each of the predefined dimensions to omit items with insufficient item characteristics. Then exploratory factor analysis guided the construction of the three-dimensional MIRT model that provided the best solution for our data.

This solution also makes sense in theoretical terms. The first factor is the predefined dimension ''severity'' (Fig. 1), containing items about the intensity, frequency, and duration of fatigue. The second factor embraces several relatively large dimensions, namely ''physical,'' ''mental,'' ''consequences,'' and ''coping.'' All items have in common that they are about the impact of fatigue, on physical and mental level and impact directly related to different aspects of daily life as already included in our predefined dimension called ''consequences.'' That the dimension ''coping'' also belongs to the second factor is of special interest. It is one of the relatively small dimensions, we did not summarize to a larger dimension for the analysis because it contains items that are not frequently included in other fatigue instruments yet. Items about what people did or did not do to cope with their fatigue can also be regarded as a consequence or impact of fatigue because these behaviors are resulting from the fatigue. The two other ''new'' dimensions, namely ''change'' and ''perceived causes'' form the third factor. They refer to the changing character described by patients [7] and the reasons patients attribute

to their fatigue. The third factor is clearly less stable than the first two in psychometric terms. However, it reflects important aspects of the patient perspective on fatigue. For a valid measurement of fatigue, it is important to find a good balance between both perspectives; psychometric results and information gained from patient experience. Modern psychometric methods as IRT include the danger of losing face validity of items and the danger of excluding items although they are needed for an adequate reflection of the measured construct [47]. To ensure that items from this third factor will be drawn in the adaptive testing process, it could be a possibility to place accordant restrictions on the CAT [48]. However, compared with unidimensional IRT, our MIRT approach provides a greater flexibility for item selection. Owing to the existence of more than one dimension, fewer items have to be excluded because they do not comply with the unidimensionality criterion. This means that the protection of content validity in our study is superior compared with unidimensional IRT.

The results of this study clearly underline the multidimensionality of fatigue as reported by patients. In exploratory factor analysis, the one-dimensional model had to be rejected. Also when comparing the three-factor model with the one-dimensional model, the multidimensional model turned out predominant. For multidimensional fatigue assessment, single-item instruments, such as frequently used visual analogue scales, are not appropriate. Adequate measures of fatigue are essential for science and clinical practice to get more insight into fatigue and its causes and impact and to be able to develop and evaluate interventions or treatments to reduce fatigue [17].

Apart from aspects regarding content, the technique of multidimensional adaptive testing has several advantages. It provides information about the level of a participant on each dimension and about the amount of association between dimensions in the population [49]. The cross-information gained from items of correlated dimensions facilitates CAT by supporting the selection of optimal, informative items, and by supporting the estimation of fatigue with optimal precision. Multidimensional adaptive testing offers equal or even higher precision with approximately one-third fewer items than would be needed in unidimensional adaptive testing [49]. With this innovative method, measuring fatigue in RA can become more precise and at the same time more user-friendly.

To our knowledge, we developed the first multidimensional fatigue item bank by applying a MIRT model to the data. Existing fatigue item banks are unidimensional [50] or when measuring with more than one dimension, one item bank per dimension was calibrated separately [35]. Until now, only few multidimensional item banks were developed to measure PROs in health care. However, studies have already demonstrated that the application of MIRT models can lead to precise and efficient multidimensional CAT in this area, for example, for dyspnea assessment [51] or to measure health-related quality of life [52].

The strength of our item pool was the thorough development from the patients' perspective and inclusion of established aspects of fatigue and also new aspects that were brought up by patients. A limitation is the relatively small sample compared with other samples used for calibration studies [26]. Further research has to show how robust the results of this study are. Possibly, the third dimension (variability of fatigue) will work out better in an analysis with more data. This study was a first, explorative approach to form the basis for the development of a multidimensional CAT for fatigue in RA, and the initial calibrated item pool will undergo further statistical examination in the future process of the CAT development. With this innovative measurement approach, it will be possible to measure fatigue in patients with RA more precisely and with fewer items [49].

Acknowledgments

The authors would like to thank the participants in this study: patients from Ziekenhuisgroep Twente (Almelo and Hengelo) and Arthritis Center Twente at Medical Spectrum Twente. Furthermore, they also thank the rheumatologists and nurses for supporting the recruitment of participants.

Appendix B. Supplementary data

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.jclinepi.2013.05.010.

References

[1] Wolfe F, Hawley DJ, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol 1996;23:1407−17.
[2] Hewlett S, Carr M, Ryan S, Kirwan J, Richards P, Carr A, et al. Outcomes generated by patients with rheumatoid arthritis: how important are they? Musculoskeletal Care 2005;3:131−42.
[3] Hewlett S, Chalder T, Choy E, Cramp F, Davis B, Dures E, et al. Fatigue in rheumatoid arthritis: time for a conceptual model. Rheumatology 2011;50(6):1004−6. http://dx.doi.org/10.1093/rheumatology/keq282.
[4] Belza Tack B. Fatigue in rheumatoid arthritis. Conditions, strategies, and consequences. Arthritis Care Res 1990;3(2):65−70.
[5] Hewlett S, Cockshott Z, Byron M, Kitchen K, Tipler S, Pope D, et al. Patients' perceptions of fatigue in rheumatoid arthritis: overwhelming, uncontrollable, ignored. Arthritis Care Res 2005;53(5):697−702.
[6] Repping-Wuts H, Uitterhoeve R, van Riel P, van Achterberg T. Fatigue as experienced by patients with rheumatoid arthritis (RA): a qualitative study. Int J Nurs Stud 2008;45:995−1002.
[7] Nikolaus S, Bode C, Taal E, van de Laar MAFJ. New insights into the experience of fatigue among patients with rheumatoid arthritis: a qualitative study. Ann Rheum Dis 2010;69(5):895−7.
[8] Repping-Wuts H, van Riel P, van Achterberg T. Fatigue in patients with rheumatoid arthritis: what is known and what is needed. Rheumatology 2009;48:207−9.
[9] Minnock P, Kirwan J, Bresnihan B. Fatigue is a reliable, sensitive and unique outcome measure in rheumatoid arthritis. Rheumatology 2009;48:1533−6.

[10] Chalder T, Berelowitz G, Pawlikowska T, Watts L, Wessely S, Wright D, et al. Development of a fatigue scale. J Psychosom Res 1993;37:147−53.
[11] Smets E, Garssen B, Bonke B, de Haes JCJM. The Multidimensional Fatigue Inventory (MFI): psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39:315−25.
[12] 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.
[13] Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. J Rheumatol 2005;32:811−9.
[14] McNair D, Lorr M, Droppelman L. Profile of mood states manual. New York, NY: Multi-health Systems, Inc.; 1992.
[15] Vercoulen J, Swanink C, Fennis J, Galama JM, van der Meer JW, Bleijenberg G. Dimensional assessment of chronic fatigue syndrome. J Psychosom Res 1994;38:383−92.
[16] Hewlett S, Hehir M, Kirwan JR. Measuring fatigue in rheumatoid arthritis: a systematic review of scales in use. Arthritis Rheum 2007;57:429−39.
[17] Hewlett S, Dures E, Almeida C. Measures of fatigue. Arthritis Care Res 2011;63:S263−86.
[18] Belza Tack B. Dimensions and correlates of fatigue in older adults with rheumatoid arthritis (dissertation). San Francisco, CA: University of California; 1991.
[19] Streiner DL, Norman GR. Health measurement scales—a practical guide to their development and use. New York, NY: Oxford University Press; 2003.
[20] Fayers PM, Machin D. Quality of life—assessment, analysis and interpretation. Chichester, UK: Wiley; 2000.
[21] Yorkston KM, Johnson K, Boesflug E, Skala J, Amtmann D. Communication about the experience of pain and fatigue in disability. Qual Life Res 2010;19:243−51.
[22] Nicklin J, Cramp F, Kirwan J, Greenwood R, Urban M, Hewlett S. Measuring fatigue in rheumatoid arthritis: a cross-sectional study to evaluate the Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional questionnaire, visual analogue scales, and numerical rating scales. Arthritis Care Res 2010;62(11):1559−68.
[23] Nicklin J, Cramp F, Kirwan J, Urban M, Hewlett S. Collaboration with patients in the design of patient reported outcome measures: capturing the experience of fatigue in rheumatoid arthritis. Arthritis Care Res 2010;62:1552−8.
[24] Rose M, Bezjak A. Logistics of collecting patient-reported outcomes (PROs) in clinical practice: an overview and practical examples. Qual Life Res 2009;18:125−36.
[25] Walter OB. Adaptive tests for measuring anxiety and depression. In: van der Linden WJ, Glas CAW, editors. Elements of adaptive testing. New York, NY: Springer; 2010:123−36.
[26] Fliege H, Becker J, Walter OB, Rose M, Bjorner JB, Klapp BF. Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. Int J Methods Psychiatr Res 2009;18(1):23−36.
[27] Becker J, Fliege H, Kocalevent RD, Bjorner JB, Rose M, Walter OB, et al. Functioning and validity of a computerized adaptive test to measure anxiety (A-CAT). Depress Anxiety 2008;25:182−94.
[28] Kocalevent RD, Rose M, Becker J, Walter OB, Fliege H, Bjorner JB, et al. An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. J Clin Epidemiol 2009;62:278−87.
[29] Wainer H. Computerized adaptive testing: a primer. Hillsdale, NJ: Lawrence Erlbaum Associate; 1990.
[30] Nikolaus S, Bode C, Taal E, van de Laar MAFJ. Four different patterns of fatigue in rheumatoid arthritis patients: results of a Q-sort study. Rheumatology 2010;49(11):2191−9.
[31] Nikolaus S, Bode C, Taal E, van de Laar MAFJ. Selection of items for a computer-adaptive test to measure fatigue in patients with rheumatoid arthritis—a Delphi approach. Qual Life Res 2012;21:863−72.
[32] Nikolaus S, Bode C, Taal E, van de Laar MAFJ. Which dimensions of fatigue should be measured in patients with rheumatoid arthritis?—a Delphi study. Musculoskeletal Care 2012;10(1):13−7.
[33] Nikolaus S, Bode C, Taal E, van de Laar MAFJ. Experts' evaluations of fatigue questionnaires used in rheumatoid arthritis—a Delphi study among patients, nurses and rheumatologists in the Netherlands. Clin Exp Rheumatol 2012;30:79−84.
[34] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH Roadmap Cooperative Group during its first two years. Med Care 2007;45:I3−11.
[35] Cella D, Lai J-S, Stone A. Self-reported fatigue: one dimension or more? Lessons from the Functional Assessment of Chronic Illness Therapy—fatigue (FACIT-F) questionnaire. Support Care Cancer 2011;19:1441−50.
[36] Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417−32.
[37] Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II. Arthritis Rheum 2004;50:3296−305.
[38] Reckase MD. Linking and scaling. In: Reckase MD, editor. Multidimensional item response theory. Statistics for Social and Behavioral Sciences. New York, NY: Springer; 2009:275−310.
[39] Glas CAW. Software program multidimensional item response theory (MIRT). 2010. Available at http://www.utwente.nl/gw/omd/afdeling/temp_test/mirt-manual.pdf.
[40] Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum; 1980.
[41] Muraki E. A generalized partial credit model: application of an EM algorithm. Appl Psychol Meas 1992;16:159−76.
[42] Embretson SE, Reise SP. Polytomous IRT models. In: Item response theory for psychologists. London, UK: Lawrence Erlbaum Associates, Publishers; 2000:95−124.
[43] Glas CAW. Modification indices for the 2-PL and the nominal response model. Psychometrika 1999;64:273−94.
[44] Van Groen MM, ten Klooster PM, Taal E, van de Laar MAFJ, Glas CAW. Applications of the health assessment questionnaire disability index to various rheumatic diseases. Qual Life Res 2010;19:1255−63.
[45] Hambleton RK, Swainathan H, Rogers HJ. Identification of potentially biased test items. In: Fundamentals of item response theory. Newbury Park, CA: Sage Publications; 1991.
[46] Thomas N, Gan N. Generating multiple imputations for matrix sampling data analyzed with item response models. J Educ Behav Stat 1997;22:425−45.
[47] Fries JF. New instruments for assessing disability: not quite ready for prime time. Arthritis Rheum 2004;50:3064−7.
[48] Van der Linden WJ. Constrained adaptive testing with shadow tests. In: van der Linden WJ, Glas CAW, editors. Elements of adaptive testing. New York: Springer; 2010:31−56.
[49] Segall DO. Principles of multidimensional adaptive testing. In: van der Linden WJ, Glas CAW, editors. Elements of adaptive testing. Statistics for Social and Behavioral Sciences. New York, NY: Springer; 2010:57−75.
[50] Lai J-S, Cella D, Dineen K, Bode R, von Roenn J, Gershon RC, et al. An item bank was created to improve the measurement of cancer-related fatigue. J Clin Epidemiol 2005;58:190−7.
[51] Norweg A, Ni P, Garshick E, O'Connor G, Wilke K, Jette AM. A multidimensional computer adaptive test approach to dyspnea assessment. Arch Phys Med Rehabil 2011;92:1561−9.
[52] Petersen MA, Groenvold M, Aaronson N, Fayers P, Sprangers M, Bjorner JB. Multidimensional computerized adaptive testing of the EORTC QLQ-C30: basic developments and evaluations. Qual Life Res 2006;15:315−29.

Appendix A

Table A1: Item administration design

Booklet	Severity: 19 items	Physical: 37 items	Mental: 54 items	Consequences I: 43 items	Consequences II: 42 items	Change: 9 items	Perceived causes I: 9 items	Perceived causes II: 9 items	Coping I: 12 items	Coping II: 11 items	Items	*N*
1	■	■				■	■	■	■	■	106	80
2	■		■	■		■	■		■		103	80
3	■			■		■	■	■	■	■	112	79
4		■	■			■	■		■		121	85
5		■			■	■		■		■	108	77
6			■	■	■	■		■		■	126	81
7	■			■	■	■					113	69
Larger dimension	1	2	3	4		5	6		7		245	551
Number excluded items (step 1)	6	6	2	10		5	8		12		49	

Columns refer to the item clusters. The shaded areas refer to administered items. The blank cells refer to not administered items. The last row refers to items removed owing to misfit or unreliability.