RHEUMATOLOGY ADVANCES IN PRACTICE

Rheumatology Advances in Practice 2020;0:1–8
doi:10.1093/rap/rkaa012

Original Article

Two parallel short forms to measure disease- and treatment-associated knowledge in rheumatoid arthritis: application of item response theory

Marieke J. Spijk-de Jonge[1], Martijn A. H. Oude Voshaar[2], Lisanne Renskers[1], Anita M. P. Huis[1], Mart A. F. J. van de Laar[2], Marlies E. J. L. Hulscher[1] and Piet L. C. M. van Riel[1]

Abstract

Objective. The aim was to develop two disease- and treatment-related knowledge about RA (DataK-RA) short forms using item response theory-based linear optimal test design.

Methods. We used the open source Excel add-in solver to program a linear optimization algorithm to develop two short forms from the DataK-RA item bank. The algorithm was instructed to optimize precision (i.e. reliability) of the scores for both short forms, subject to a number of constraints that served to ensure that each short form would include unique items and that the short forms would have similar psychometric properties. Agreement among item response theory scores obtained from the different short forms was assessed using the Bland–Altman method and Student's paired *t*-test. Construct validity and relative efficiency of the short forms was evaluated by relating the score to age, sex and educational attainment.

Results. Two short forms were derived from the DataK-RA item bank that satisfied all content constraints. Both short forms included 15 unique items and yielded reliable scores ($r > 0.70$), with low ceiling and floor effects. The short forms yielded statistically indistinguishable mean scores according to Student's paired *t*-test and Bland–Altman analysis. Scores on short forms 1 and 2 were associated with age, sex and educational attainment to a similar extent.

Conclusion. In this study, we developed two DataK-RA short forms with unique items, yet similar psychometric properties, that can be used to assess patients pre- and post-test interventions aimed at improving disease-related knowledge in RA patients.

Key words: item response theory, patient education, patient knowledge, rheumatoid arthritis, short form

Key messages

- We developed disease- and treatment-related knowledge about RA short forms for assessing patients' knowledge with minimal patient burden.
- Both short forms yield reliable scores to measure patient knowledge levels in pre–post-test intervention studies.
- Item response theory enables item selection while still allowing comparison of outcomes with earlier studies.

[1]Radboud University Medical Center, Radboud Institute for Health Sciences, IQ Healthcare, Nijmegen and [2]Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands

Submitted 28 January 2020; accepted 16 April 2020

Correspondence to: Martijn A. H. Oude Voshaar, Department of Psychology, Health and Technology, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands.
E-mail: a.h.oudevoshaar@utwente.nl

Introduction

Shared decision-making is a collaborative process engaged in by patients and their health-care providers, in which patient values and preferences, in addition to medical evidence, are taken into account when making treatment decisions that are optimally personalized to the circumstances, needs and preferences of individual patients [1, 2]. Knowledge about their disease and its treatment is a prerequisite for patients to engage fully in shared decision-making [3]. Patient education interventions have, therefore, been recommended as an integral part of the management of patients with RA [4–6]. It is also recommended that health-care professionals make use of evidence-based education programmes to educate patients about their disease. Patient knowledge tests are used in clinical trials to assess the increase in disease knowledge attributable to education interventions.

In a previous paper, we introduced the disease- and treatment-associated knowledge in rheumatoid arthritis (DataK-RA) item bank for assessing patients' overall level of disease knowledge. DataK-RA comprehensively captures different aspects of knowledge of RA and its treatment [7]. The item bank was developed using an extensive research process, in which content of previously validated knowledge questionnaires was combined with contemporary treatment insights of health-care professionals and patients [8–11].

Although most previously proposed patient knowledge instruments used in RA are based on classical test theory, we chose to develop an item response theory (IRT)-based item bank. With respect to the assessment of patient knowledge in particular, IRT has a number of advantages. First, IRT allows users of the DataK-RA item bank to select only those items from the item bank that they consider most relevant for their patient population, while still allowing the results of their study to be compared with the results of other studies for which different items were chosen. Second, different items may also be presented to the same patient at different time points. This is particularly useful in the assessment of patient knowledge in pre–post-test intervention studies, because presenting patients with the same questionnaire at multiple time points may result in an overestimation of the increase in knowledge, as a result of patients remembering or learning the answers to specific questions. Finally, which information about their disease is important for patients to know may change over time as new insights into RA or its treatment develop, causing certain items to lose relevance and new information to become more important. Item response theory allows obsolete items to be removed and new items to be added to the item bank, without losing the ability to compare outcomes with earlier studies.

However, a potential drawback of IRT-based instruments is that scoring procedures are fairly complex and usually require specific software to implement, which might limit the viability of IRT-based instruments in practice. Therefore, in the present paper we introduce two DataK-RA-derived questionnaires that can be used by health-care professionals or researchers aiming to assess patients' knowledge levels with minimal burden. We provide an evaluation of the psychometric properties of both short forms and provide instructions on how to obtain DataK-RA IRT scores for the short forms.

Methods

DataK-RA

The DataK-RA item bank encompasses items to measure comprehensively patients' knowledge on relevant aspects of RA and its treatment. These aspects were identified in a rigorous qualitative process that included a systematic literature review, a RAND-modified Delphi scoring procedure and consensus meeting with rheumatology professionals and a focus group with patients with RA. The complete item bank can be found in the supplemental materials of our earlier paper describing the development process of DataK-RA in detail [7]. The item bank contains 42 multiple choice items, with two to four response options per item.

DataK-RA was calibrated using the two-parameter logistic IRT model (2pl), which is an item response model for dichotomous items, in which the patient knowledge scores and the difficulty of the items are placed on a common scale. In particular, the model describes the probability that a person correctly answers an item as a logistic function of the patient's knowledge level (θ) minus the difficulty of the item (β). In addition, each item has another parameter (α), which determines the slope of the function. Scores on items with a high α parameter depend strongly on the latent variable, hence these items discriminate well between patients with high and low levels of disease knowledge. The item parameters can also be used to calculate item information functions that describe the contribution of the item to the precision of the scores. Summing of the item information functions of all items that were administered to a patient yields a score information function, which is inversely related to the standard error of estimation for a particular score.

Development of short forms

Cross-sectional data that were collected for the development of the DataK-RA were also available to evaluate the two short forms. These data were obtained by sending a questionnaire containing the initial pool of 63 items (in Dutch) to all 721 patients with RA from Bernhoven Rheumatology Department and all 90 patients with RA from the Rheumatology Research Panel of the University of Twente. All patients received a questionnaire via mail and received one reminder if necessary. Based on these data, the final DataK-RA item pool (42 items) was compiled [7].

This study was performed in compliance with the Declaration of Helsinki. The Committee on Research

Involving Human Subjects Arnhem–Nijmegen exempted our study from formal ethical approval because it did not involve research covered by the Medical Research Involving Human Subjects Act (file 2015-1728). In addition, the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences of the University of Twente approved our study.

In our present study, we used the mixed integer programming method proposed by Van der Linden [12] to derive two DataK-RA short forms with optimal and similar measurement properties. We wanted the short forms to be useful in longitudinal studies in RA patients who have not yet been exposed to educational interventions aimed at increasing RA-related knowledge. Therefore, we sought to include items in the short forms that were optimally suited to assess disease knowledge levels of RA patients in the subpopulation of patients with a relatively low level of knowledge. Also, we wanted each short form to include 15 items to prevent patient burden, and each item could feature in only one short form to prevent learning effects. Furthermore, the short forms should yield reliable scores of similar precision across the continuum of patient knowledge scores.

These requirements can be framed as a combinatorial optimization problem, where the objective function is to maximize the scale information functions of the two short forms given certain content constraints imposed on the included items. We used the open source Excel add-in solver to program the optimization algorithm. Given that the scores in the overall sample of patients were normally distributed with a mean (s.d.) IRT score of 50 (10), we chose to optimize the information functions at the three IRT score levels of 30, 40 and 50, subject to the constraints that: (1) the total number of items included in each short form should be 15; (2) the conditional reliability (*r*) coefficients should be $\geq$0.70 at each of the three IRT score levels of 30, 40 and 50 [13]; (3) the absolute difference in information provided by the short forms should not exceed 0.50 at any of the three IRT score levels; and (4) each item can feature in only one short form.

Obtaining short form scores

We created several tools that researchers can use to obtain IRT-scaled scores for response data collected using either or both short forms. Initially, we tabulated expected *a posteriori* scores and associated standard errors for each possible raw score, for both short forms. The resulting crosswalk table assigns the same IRT-scaled score to all response patterns that lead to the same number of correct responses. However, the conversion tables are applicable only when there are no missing values. Moreover, it is likely that IRT-scaled scores for individual response patterns are slightly more accurate and precise compared with the IRT-scaled scores obtained using the conversion tables. Therefore, we compared the overall reliability and agreement of the crosswalked scores with expected *a posteriori* scores for individual response patterns. R code to obtain IRT-scaled scores when missing values are present or when optimal accuracy is sought is available from the corresponding author on request.

Score agreement

The agreement among IRT scores obtained from both short forms was examined using the Bland–Altman procedure [17]. Given that the IRT model serves to correct the scores for item characteristics, we expected that the estimated bias should not be significantly different from zero. This was tested using a one-sample Student's *t*-test. We also compared agreement among scores obtained using different short forms between crosswalked IRT-scaled and IRT pattern scores.

Reliability and measurement precision

The reliability of both short forms was tested using greatest empirical reliability coefficients [15]. A reliability coefficient $>$0.70 has been proposed as a minimum standard for use in scientific studies for group-level inferences [16]. To examine the degree to which short form scores were equally precise across the different patient knowledge levels, we plotted the information functions for both short forms and compared these visually.

Construct validity and relative efficiency

In our previous paper, DataK-RA total scores were found to increase with educational attainment according to the International Standard Classification of Education level and age, and female patients were found to have higher DataK-RA total scores than men [7]. In the present study, we therefore hypothesized that DataK-RA scores on both short forms would also be associated with these variables and that the strength of these associations would be similar for both short forms. These hypotheses were tested using univariate linear regression analysis. For each variable, we obtained the proportion of explained variance (R^2), and we tested the hypothesis that the slope of the regression line was statistically different from zero, using a Student's *t*-test provided by SPSS version 23.

In addition, given that both short forms assess the same construct and were developed to have similar reliability, we hypothesized that the relative efficiency, defined as the ratio of the test statistics for the regression coefficients, would be close to one for age, educational level and sex [17].

Floor and ceiling effects

Floor and ceiling effects, defined as the proportions of patients with 0 and 100% correct answers, respectively, were examined and compared between the different short forms. Floor or ceiling effects $>$15% are usually considered problematic [18]. We tested the difference between the short forms using the *T*-test for dependent proportions.

Results

Patient characteristics

Of the 811 patients with RA who received the questionnaire, 419 patients recruited from Bernhoven and 54 patients recruited from the University of Twente returned a completed questionnaire. This corresponds to a response rate of 58 and 60%, respectively. The characteristics of these patients represented a typical RA population, with more females than males (64.5%), and the mean age was 65 years (s.d. = 13 years). Average disease duration in our sample was 13 months (s.d. = 12; see Table 1). Generally, patients filled in the complete questionnaire. The mean percentage of missing values for items was 3.2% (s.d. = 2.0%).

Short forms

Two short forms could be derived from the DataK-RA item bank that satisfied the content constraints. The items and response options of both short forms are presented in the Supplemental Material, available at *Rheumatology Advances in Practice* online. Table 2 presents the conversions that can be used to obtain approximate DataK-RA IRT scores from the summed scores for both short forms. These conversion tables can be used to convert the raw number correct scores to IRT-scaled scores. Table 2 shows that IRT-scaled scores can range from ~24 to 65 for both short forms and that raw scores on both short forms were linked to similar IRT-scaled scores. Given that the scale information functions were constrained to be similar (constraint 3 in the Methods section), conditional reliability coefficients associated with each raw score were similar across the range of raw scores from 0 to 15. Finally, it can be seen that conditional reliability was >0.70 across the range of IRT-scaled scores from 30 to 50 (constraint 2) for both short forms.

Score reliability

Item response theory-scaled scores were reliable, with all reliability coefficients >0.70. The reliability coefficients for IRT pattern scores were slightly higher (0.71, 0.73) compared with reliability coefficients obtained from IRT-scaled scores converted using Table 2 (0.70, 0.72) for short form 1 and 2, respectively. As intended, the precision of the scores was maximum at the targeted IRT score levels (Fig. 1), and the precision of short form scores was similar across the latent patient knowledge continuum.

Table 1 Demographics of the respondents ($n = 473$)

Parameter	Value
Sex, *n* (% female)	305 (64.50%)
Age, years	64.99 (13.00) (23–101)
Disease duration, months	12.96 (11.99) (1–77)
Education level, *n* (%)	
Low	199 (42.8)
Intermediate	197 (42.5)
High	68 (14.4)

Values are the mean (s.d.) (range), unless indicated otherwise.

Score agreement

The IRT pattern scores on short form 1 exhibited an approximately normal distribution and those short form 2 exhibited an almost identical distribution, with a mean (s.d.) of 51.06 (8.39) on short form 1 and 51.28 (8.39) on short form 2. The IRT pattern scores were slightly lower on short form 1, with an estimated bias (s.d.) of −0.218 (5.88). However, this was not significantly different from 0 (t-distribution = 0.68, $P = 0.49$). The 95% agreement interval ranged from −11.75 to 11.31 points on the *t*-score metric (Fig. 2). Taken together, these results indicate that there is no systematic bias in scores obtained from the two short forms when the IRT-based pattern scoring procedure is used and that scores on the short forms can be expected to lie within a range of 11 points of each other.

The approximate IRT scores obtained via the crosswalk were only slightly less congruent, with an estimated bias (s.d.) of −0.257 (5.88), ($P = 0.42$) and limits of agreement ranging from −11.77 to 11.25. This suggests that when there are no missing values, researchers can use the conversion tables, without loss of accuracy and precision of the scores.

Construct validity and relative efficiency

The results of the linear regression analyses are summarized in Table 3. The results were generally in agreement with our hypotheses. Age, educational level and sex were all associated with DataK-RA short form score, except that the scores on short form 1 were not statistically associated with sex. In addition, the relative efficiency coefficients were all close to one, which suggests that the strength of the associations of DataK-RA scores with the tested variables was similar for short forms 1 and 2.

Floor and ceiling effects

There were no floor effects for either short form, with no (0%) patients with no correct answers on short form 1 and one (0.2%) person with no questions right on short form 2. Floor effects were similar for both questionnaires ($z = 0.97$, $P = 0.16$). There were also no notable ceiling effects, with 37 (7.8%) and 31 (6.6%) patients with all questions correct on short forms 1 and 2, respectively. The percentage of patients with all questions correct did not differ significantly ($z = 0.68$, $P = 0.25$).

Discussion

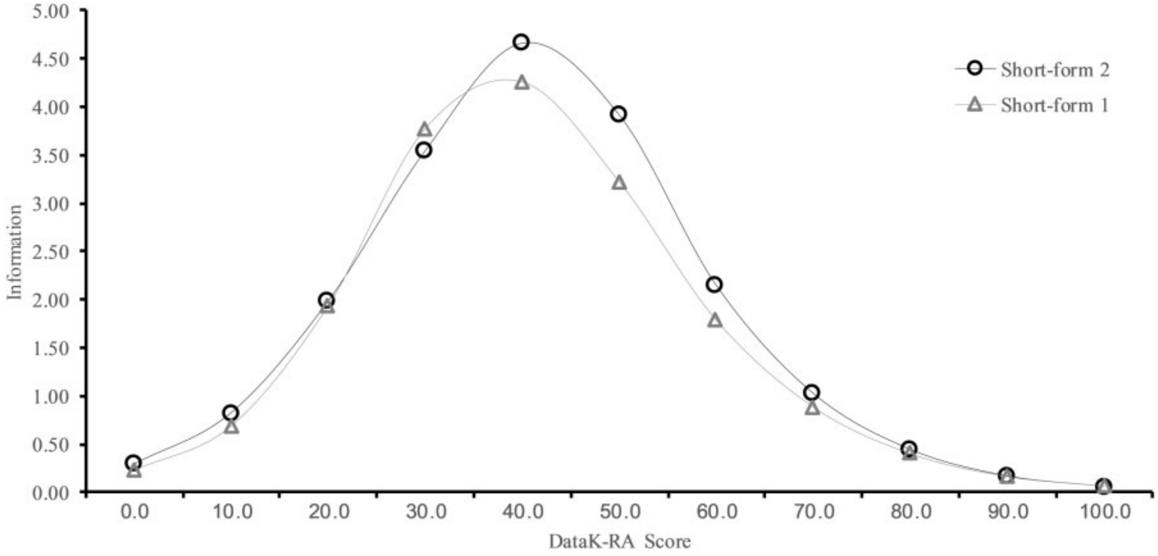
In this paper, we introduced two short forms based on the DataK-RA item bank. Items from DataK-RA can be

Table 2 Crosswalks

Raw score	short form 1, *t*-score	Short form 1, s.e.	CR	Short form 2, *t*-score	Short form 2, s.e.	CR
0	23.75	5.53	0.69	23.87	5.42	0.71
1	26.88	5.13	0.74	27.10	4.98	0.75
2	29.70	4.81	0.77	29.94	4.67	0.78
3	32.27	4.57	0.79	32.51	4.44	0.80
4	34.68	4.40	0.81	34.89	4.29	0.82
5	36.97	4.30	0.82	37.14	4.19	0.82
6	39.21	4.26	0.82	39.30	4.15	0.83
7	41.43	4.27	0.82	41.44	4.15	0.83
8	43.68	4.34	0.81	43.60	4.19	0.82
9	46.03	4.47	0.80	45.83	4.28	0.82
10	48.51	4.66	0.78	48.20	4.43	0.80
11	51.20	4.93	0.76	50.76	4.65	0.78
12	54.17	5.26	0.72	53.60	4.94	0.76
13	57.47	5.67	0.68	56.85	5.34	0.72
14	61.18	6.14	0.62	60.65	5.85	0.66
15	65.46	6.68	0.55	65.31	6.53	0.57

CR: conditional reliability.

Fig. 1 Score precision



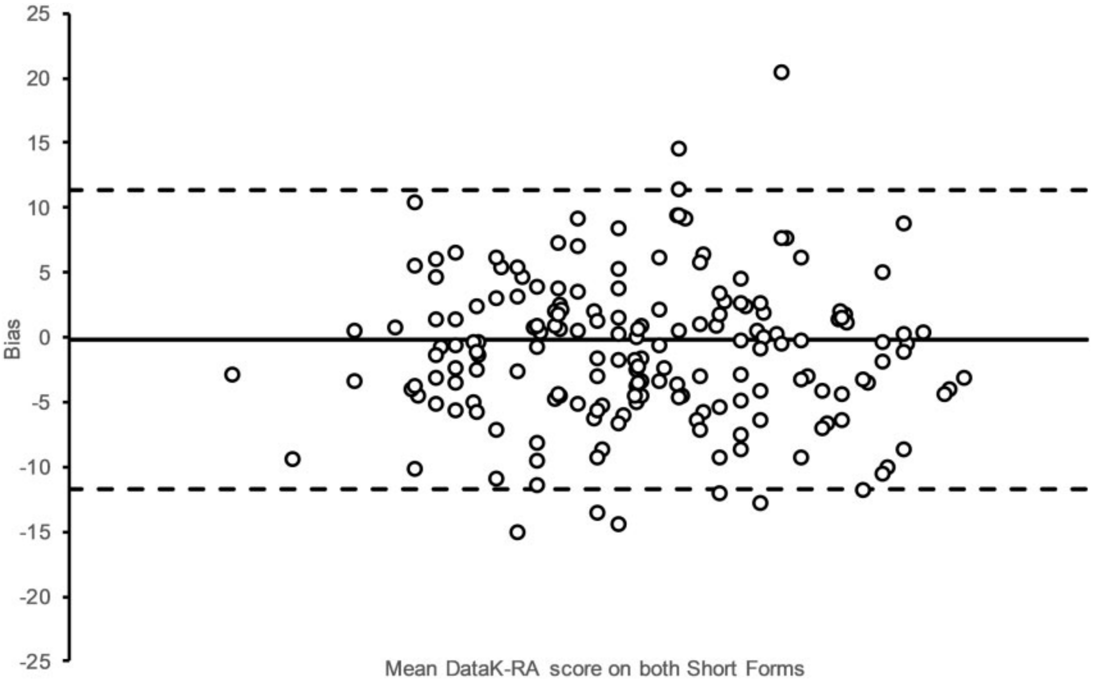
Datak-RA: disease- and treatment-related knowledge about RA.

used to measure RA patients' knowledge and to identify possible gaps in their knowledge to target patient education. This allows users to adapt educational interventions to the needs of patients. Moreover, it can help to assess the effects of educational interventions on the knowledge levels of (groups of) patients. We developed DataK-RA based short forms that can be used to measure patients' knowledge level with minimal patient burden.

We also provided several tools that researchers interested in using these short forms can use to obtain IRT-scaled DataK-RA scores for their own data collected with either or both of these short forms. The results of the present study show that the crosswalked IRT-scaled scores performed in a similar manner to the IRT pattern scores in terms of agreement among alternative versions and overall reliability of the scores. Based on these findings, we conclude that researchers interested in using the short forms can confidently use the crosswalked scores, without great loss of accuracy and precision.

Moreover, given that each short form included unique items, they can be used by researchers interested in assessing patient knowledge of RA as a means to control item exposure and to remove bias attributable to learning effects from their studies. This makes them suitable for use in pre–post-intervention studies. Given that the items in both short forms do not overlap, the

Fig. 2 Item response theory-based pattern scoring



Bias is the score on short form 1 minus the score on short form 2. Datak-RA: disease- and treatment-related knowledge about RA.

Table 3 Construct validity and relative efficiency

	Short-form 1			Short form 2			
	β (constant)	R^2	*t*	β (constant)	R^2	*t*	RE
Age, years	−0.37 (78.20)	0.15	−7.85*	−0.30 (69.56)	0.16	−8.14*	0.96
Education	9.26 (37.38)	0.29	11.45*	5.77 (40.00)	0.18	8.51*	0.74
Sex[a]	−2.77 (58.33)	<0.01	−1.80	−2.29 (53.70)	0.01	−1.94*	0.93

Educational attainment levels in accordance with the International Standard Classification of Education: 1=low; 2=intermediate; 3=high. [a]1= female, 2= male. *Statistically significant at the 0.05 level. T: T-test statistic; RE: relative efficiency; R^2: proportion of explained variance; *t*: Student's *t*-test for the slope of the regression line.

improvement in score over the course of study cannot be attributed to patients having learned the correct answers to individual questions. In support of the construct validity, both short forms were found to relate to age, education level and sex to a similar extent. Similar associations have been found for other patient knowledge questionnaires in previous studies and in our own analysis of the full DataK-RA item bank [7–11]. Our results also showed that both developed short forms had good psychometric properties when comparing the reliability and floor and ceiling effects with common benchmark or cut-off values for high-quality measurement properties, with reliability coefficients >0.70 and floor/ceiling effects <15% [14]. Furthermore, the IRT analysis of the information functions showed that reliable scores can be obtained across the spectrum, ranging from extremely low patient knowledge (2 s.d. below the mean) to ∼0.5 s.d. above the mean. These findings indicate that higher reliability and lower ceiling effects will be obtained if the questionnaires are applied in populations with lower disease knowledge than our sample of patients. On the contrary, the instruments are not as well suited for measuring higher levels of disease knowledge. This is a direct result of the choices we made in the item selection procedure. For use in populations with higher or unknown levels of disease knowledge, different item selections or computerized adaptive testing will yield better results.

Our results also showed a high level of agreement among IRT scores obtained for the two short forms, irrespective of whether IRT-based pattern scoring or the crosswalks were used for scoring. This was an expected finding for two reasons. The first reason is that the item response model corrects the IRT scores for item characteristics, such as the difficulty of the items. The second reason is that the optimal test assembly algorithm was programmed to ensure similar statistical information for both short forms. As is likely to happen, in our case this has led to a balanced distribution of the items over the short forms in terms of their difficulty parameters, meaning that even the expected summed scores are more similar for the different forms than would be the case if the items were randomly distributed between the two forms.

Although the high agreement of IRT scores for both versions and high reliability suggest that the short forms should be responsive to change in patient knowledge, in the present study design we were not able to evaluate this. Another potential limitation is that we used only psychometric performance criteria for the item selection process. Short forms balanced with respect to specific item content, number of response options or other criteria can, in principle, also be derived. However, the current version of the item bank contains only 42 items, which limits the potential for such applications. Finally, there was a relatively low response rate. The representativeness of these results for the overall patient populations in our clinical setting is therefore unclear. Future research should be directed at including more items in the item bank, in particular more difficult items, which would additionally increase the measurement performance of DataK-RA-derived measures in populations with higher levels of disease-related knowledge.

Our ongoing research activities are aimed to develop the DataK-RA and short forms further on these points. Currently, the short forms described in this paper are being used in an intervention study, meaning that we will be able to evaluate the sensitivity to change of DataK-RA. Also, we are working on the development of additional items for the item bank, specifically focusing on the inclusion of more difficult items and items on recent developments in RA treatment. We are also working on validation of the English version of DataK-RA.

In sum, the development of DataK-RA short forms is yet another step in providing health-care professionals and researchers with psychometrically sound and up-to-date tools to assess disease-related knowledge in RA patients. We hope that these short forms prove useful in targeted patient education and in measuring whether education improves knowledge.

Acknowledgements

We would like to thank all rheumatology nurses, rheumatologists and patients who have helped in the development or validation of the item bank and its short forms.

Funding: This work was supported by the Dutch Arthritis Association [project BP12-1-262].

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

Supplementary data

Supplementary data are available at *Rheumatology Advances in Practice* online.

References

1 Elwyn G, Tilburt J, Montori V. The ethical imperative for shared decision-making. Eur J Pers Cent Healthc 2013; 1:129–31.

2 Voshaar MJ, Nota I, van de Laar MA, van den Bemt BJ. Patient-centred care in established rheumatoid arthritis. Best Pract Res Clin Rheumatol 2015;29:643–63.

3 Joseph-Williams N, Elwyn G, Edwards A. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ Couns 2014;94: 291–309.

4 de Wit MP, Smolen JS, Gossec L, van der Heijde DM. Treating rheumatoid arthritis to target: the patient version of the international recommendations. Ann Rheum Dis 2011;70:891–5.

5 van Eijk-Hustings Y, van Tubergen A, Boström C *et al*. EULAR recommendations for the role of the nurse in the management of chronic inflammatory arthritis. Ann Rheum Dis 2012;71:13–9.

6 Zangi HA, Ndosi M, Adams J *et al*. EULAR recommendations for patient education for people with inflammatory arthritis. Ann Rheum Dis 2015;74:954–62.

7 de Jonge MJ, Oude Voshaar MAH, Huis AMP *et al*. Development of an item bank to measure factual disease and treatment related knowledge of rheumatoid arthritis patients in the treat to target era. Patient Educ Couns 2018;101:67–73.

8 Edworthy SM, Devins GM, Watson MM. The arthritis knowledge questionnaire. A test for measuring patient knowledge of arthritis and its self-management. Arthritis Rheum 1995;38:590–600.

9 Hennell SL, Brownsell C, Dawson JK. Development, validation and use of a patient knowledge questionnaire (PKQ) for patients with early rheumatoid arthritis. Rheumatology 2004;43:467–71.

10 Hill J, Bird HA, Hopkins R, Lawton C, Wright V. The development and use of Patient Knowledge Questionnaire in rheumatoid arthritis. Br J Rheumatol 1991;30:45–9.

11 Lineker SC, Badley EM, Hughes EA, Bell MJ. Development of an instrument to measure knowledge in individuals with rheumatoid arthritis: the ACREU rheumatoid arthritis knowledge questionnaire. J Rheumatol 1997;24:647–53.

12 van der Linden WJ. Linear models for optimal test design. New York: Springer, 2005.

13 Raju NS, Price LR, Oshima TC, Nering ML. Standardized conditional SEM: a case for conditional reliability. Appl Psychol Meas 2007;31:169–80.

14 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307–17.

15 Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw 2012;48:1–26.

16 Nunnally JC. Psychometric theory. New York: McGraw-Hill, 1978: 2nd edn.

17 Fayers P, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. 2nd edn. Chichester: John Wiley & Sons, 2007.

18 Reeve BB, Wyrwich KW, Wu AW *et al*. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 2013;22:1889–905.