Two parallel short forms to measure disease and treatment associated knowledge in rheumatoid arthritis: Application of Item Response Theory

Marieke J. Spijk-de Jonge¹, Martijn AH. Oude Voshaar², Lisanne Renskers¹, Anita MP. Huis¹, Mart AFJ. van de Laar², Marlies EJL. Hulscher¹, Piet LCM. van Riel¹.

¹Radboud university medical center, Radboud Institute for Health Sciences, IQ healthcare, Nijmegen, The Netherlands
²University of Twente, Department of Psychology, Health and Technology, Enschede, The Netherlands

Correspondence to:
Martijn A.H. Oude Voshaar
University of Twente, Department of Psychology, Health and Technology, Enschede, The Netherlands
PO box 217, 7500 AE Enschede
The Netherlands
E-mail: a.h.ouvevoshaar@utwente.nl
ORCID iD: 0000-0003-2438-9973
Disclosure statement: the authors have declared no conflicts of interest.

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

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.
ABSTRACT

Objectives: To develop two Disease and treatment related Knowledge about Rheumatoid Arthritis (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 the short forms would have similar psychometric properties. Agreement among Item Response Theory (IRT) scores obtained from the different short forms was assessed using the Bland-Altman method and a T-test for paired samples. 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 the 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 outcomes of pre-posttest interventions aimed at improving disease related knowledge in RA patients.
KEYWORDS

- Item Response Theory
- (Patient) education
- (Patient) knowledge
- Rheumatoid arthritis
- Short form

KEY MESSAGES

- We developed DataK-RA short forms for assessing patients’ knowledge with minimal patient burden.
- Both short forms yield reliable scores to measure patient knowledge levels in pre-posttest intervention studies.
- Item Response Theory enables item selection while still allowing comparison of outcomes to earlier studies.
INTRODUCTION

Shared decision making is a collaborative process engaged in by patients and their healthcare providers, in which patient values and preferences, as well as medical evidence are taken into account when making treatment decisions that are optimally personalized to individual patients’ circumstances, needs and preferences. Knowledge about their disease and its treatment is a prerequisite for patients to fully engage in shared decision making. (1, 2) Patient education interventions have, therefore, been recommended as an integral part of the management of patients with Rheumatoid Arthritis (RA). (4-6) It is also recommended that healthcare professionals make use of evidence based education programs to educate patients about their disease. Patient knowledge tests are used in clinical trials to assess increase in disease knowledge due 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 healthcare professionals and patients. (8-11)

While 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 allows for a number of advantages. First of all, 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. Secondly, 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-posttest intervention studies, since presenting patients with the same questionnaire at multiple time points may result in an overestimation of the increase in knowledge, due to 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. IRT 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 healthcare 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 Disease and treatment Associated Knowledge in Rheumatoid Arthritis (DataK-RA) item bank encompasses items to comprehensively measure knowledge on relevant aspects of RA and its treatment. These aspects were identified in a rigorous qualitative process including 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 details.(7) The item bank contains 42 multiple choice items with 2-4 response alternatives per item.

DataK-RA was calibrated using the two parameter logistic item response theory (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, ‘α’, parameter 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 which describe the item’s contribution 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 available to also 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. 

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 of our study.

In our present study, we used the mixed integer programming method proposed by Van der Linden 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 and similarly precise scores 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. Since the scores in the overall sample of patients were normally distributed with a mean (SD) 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:

- The total number of items included in each short form should be 15;
- the conditional reliability (r) coefficients should be ≥0.70 at each of the three IRT score levels of 30, 40, and 50.(16);
- the absolute difference in information provided by the short forms should not exceed 0.50 at any of the three IRT score levels;
- each item can only feature in 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. Firstly, we tabulated expected a posteriori (EAP) 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 only applicable 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 cross-walked scores with EAP scores for individual response patterns. R-code to obtain IRT
scaled scores when missing values are present or when optimal accuracy is sought are 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) Since the IRT model serves to correct the scores for item characteristics, we expected that the estimated bias should not be significantly different from 0. This was tested using a one sample t-test. We also compared agreement among scores obtained using different short forms between cross walked IRT scaled and IRT pattern scores.

Reliability and measurement precision
The reliability of both short forms was tested using greatest empirical reliability coefficients.(18) Reliability coefficients > 0.70 has been proposed as a minimum standard for use in scientific studies for group level inferences.(19) To examine the degree to which short forms 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 0, using a T-test provided by SPSS.

In addition, since both short forms assess the same construct and were developed to be similarly reliable we hypothesized that the relative efficiency, defined as the ratio of the test statistics for the regression coefficients, would be close to 1 for age, educational level and sex.(20)

Floor and ceiling effects

Floor and ceiling effects, defined as the proportions of patients with respectively 0% and 100% correct answers were examined and compared between the different short forms. Floor or ceiling effects > 15% are usually considered problematic.(21) Again we tested the difference between the short forms using the Z-test for 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 with 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 (SD=13). Average disease duration in our sample was 13 months (SD= 12, see table 1). Generally, patients filled in the complete questionnaire. The mean percentage of missing values for items was 3.2% (SD = 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 approximately 24 to 65 for both short forms and that raw scores on both short forms were linked to similar IRT scaled scores. Because the scale information functions were constrained to be similar (constraint 3 in the method 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

IRT 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 IRT scaled scores converted using Table 2 (0.70, 0.72) for short form 1 and 2 respectively. As intended, precision of the scores was maximum at the targeted IRT score levels (figure 1) and the precision of short form scores was similar across the latent patient knowledge continuum.

**Score agreement**

The IRT pattern scores on short form 1 and short form 2 were approximately normally and nearly identically distributed, with mean (SD) 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 (SD) of -0.218 (5.88). However, this was not significantly different from 0 (T= 0.68, p =0.49). The 95% agreement interval ranged from -11.75 to 11.31 points on the T-score metric (figure 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 (SD) of -0.257 (5.88), (p = 0.42) and LOA 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 1, 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 0 (0%) patients with no correct answers on short form 1 and 1 (0.2%) person with 0 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 respectively short form 1 and short form 2. 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 Disease and treatment Associated Knowledge in Rheumatoid Arthritis (DataK-RA) item bank. Items from DataK-RA can be 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 patients’ needs. Moreover, it can help to assess the effects of educational interventions on (groups of) patients’ knowledge levels. 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 this study show that the cross walked IRT scaled scores performed similarly as the IRT pattern scores in terms of agreement among alternate 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 cross walked scores, without great loss of accuracy and precision.

Moreover, since each short form included unique items, they can be used by researchers interested in assesses patient knowledge of RA as a means to control item exposure and to remove bias due to learning effects from their studies. This makes them suitable for use in pre-post intervention studies. Since the items in both short forms do not overlap, the score improvement 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 as well our own analysis of the full DataK-RA item bank.(7-11) Our results further show 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 over 0.70 and floor/ceiling effects <15%.(17) Furthermore, the IRT analysis of the information functions shows that reliable scores can be obtained across the spectrum ranging from extremely low patient knowledge (2 SDs below the mean) to about 0.5 SDs 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 other hand, 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 items selections or computerized adaptive testing will yield better results.

Our results further show 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, this has in our case led to a balanced distribution of the items over the short forms in terms of their difficulty parameters, so that even the expected summed scores are more similar for the different forms than would be the case if the items were just 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 current study design we were not yet able to evaluate this. Another potential limitation is that we only used 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 only contains 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, particularly 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 further develop the DataK-RA and short forms on these points. Currently, the short forms described in this paper are used in an intervention study, so that we will be able to evaluate the sensitivity to change of DataK-RA. Also, we are working on developing additional items for the item bank, specifically focusing on including more difficult items and items on recent developments in RA treatment. We are also working on validating the English version of DataK-RA.

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

1. Elwyn G, Tilburt J, Montori V. The ethical imperative for shared decision-making. Eur J Pers Cent Healthc. 2013;1(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(4-5):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(3):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(6):891-5.

5. van Eijk-Hustings Y, van Tubergen A, Bostrom C, Braychenko E, Buss B, Felix J, et al. EULAR recommendations for the role of the nurse in the management of chronic inflammatory arthritis. Ann Rheum Dis. 2012;71(1):13-9.

6. Zangi HA, Ndosi M, Adams J, Andersen L, Bode C, Bostrom C, et al. EULAR recommendations for patient education for people with inflammatory arthritis. Ann Rheum Dis. 2015;74(6):954-62.

7. de Jonge MJ, Oude Voshaar MAH, Huis AMP, van de Laar M, Hulscher M, van Riel P. 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(1):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(5):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 (Oxford). 2004;43(4):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(1):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(4):647-53.

12. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. 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(11):1179-94.

13. Oude Voshaar MAH, Vonkeman HE, Courvoisier D, Finckh A, Gossec L, Leung YY, et al. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Life Res. 2019;28(1):187-97.

14. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007;57(8):1358-62.

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

16. Raju NS, Price LR, Oshima TC, Nering ML. Standardized conditional SEM: a case for conditional reliability. Appl Psychol Meas. 2007;31(2):169-180.

17. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician. 1983;32(3):307-17.

18. Chalmers RP. Mirt: A multidimensional Item Response Theory package for the R environment. J Stat Softw. 2012;48(6):1-26.
19. Nunnally JC. Psychometric Theory. 2nd ed. New York: McGraw-Hill; 1978.

20. Fayers P, Machin D. Quality of Life: The assessment, analysis and interpretation of patient-reported outcomes. 2nd ed. Chichester: John Wiley & Sons; 2007.

21. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, 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(8):1889-905.
## TABLES

### Table 1: Demographics of the respondents (N=473)

| Category                        | Value               |
|---------------------------------|---------------------|
| Gender, 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 mean ± S.D. (range), unless indicated otherwise
Table 2: Crosswalks

| Raw score | T-Score short form 1 | SE short form 1 | CR short form 1 | T-Score short form 2 | SE short form 2 | CR short form 2 |
|-----------|----------------------|-----------------|----------------|----------------------|-----------------|----------------|
| 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           |

SE= Standard Error; CR= conditional reliability
Table 3: Construct validity and relative efficiency

|                          | Short-form 1 |            | Short form 2 |            | RE |
|--------------------------|--------------|------------|--------------|------------|----|
|                          | Beta (constant) | $R^2$ | t | Beta (constant) | $R^2$ | t |           |
| **Age in 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**                  | -2.77 (58.33) | <0.01 | -1.80 | -2.29 (53.70) | 0.01 | -1.94* | 0.93 |

Educational attainment levels 1; low 2; intermediate; 3 high, in accordance with the International Standard Classification of Education; $F$= f-test, $r_{e}$= relative efficiency, $R^2$ = proportion of explained variance; $T$= t-test for the slope of the regression line; * = statistically significant at the 0.05 level; + 1= female, 2= male
FIGURES

Figure 1: Score precision

Datak-RA = Disease and Treatment related Knowledge Rheumatoid Arthritis

Figure 2: IRT based pattern scoring

Datak-RA = Disease and Treatment related Knowledge Rheumatoid Arthritis; Bias is the score on Short-form 1 minus the score on Short-form 2.
DataK-RA: disease- and treatment-related knowledge about RA.
Fig. 2: Item response theory-based pattern scoring

Mean DataK-RA score on both Short Forms