A Common Functional Disability Score for Young People with Juvenile Idiopathic Arthritis

DOI:
10.1002/acr.24204
10.1002/acr.24204

Document Version
Accepted author manuscript

DR. STEPHANIE J W SHOOP-WORRALL (Orcid ID : 0000-0002-9441-5535)

MR. MARTIJN ANTONIUS HENDRIKUS OUDE VOSHAAR (Orcid ID : 0000-0003-2438-9973)

Article type : Original Article

A Common Functional Ability Score for Young People with Juvenile Idiopathic Arthritis

AUTHORS

Stephanie JW Shoop-Worrall, PhD * (stephanie.shoop-worrall@manchester.ac.uk[1,2])

Martijn AH Oude Voshaar, PhD (a.h.oudevoshaar@utwente.nl) [3]

Janet E McDonagh MD FRCP (janet.mcdonagh@manchester.ac.uk [1,4)

Mart AFJ Van de Laar MD PhD(m.vandelaar@mst.nl) [3]

Nico Wulffraat MD PhD (N.Wulffraat@umcutrecht.nl) [5]

CAPS

Wendy Thomson PhD (wendy.thomson@manchester.ac.uk [4,6]

Kimme L Hyrich, MD PhD FRCPC (Kimme.hyrich@manchester.ac.uk [1,4])

Suzanne MM Verstappen, PhD (Suzanne.verstappen@manchester.ac.uk [1,4])

[1] Centre for Epidemiology Versus Arthritis, Division of Musculoskeletal and Dermatological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Sciences Centre, Manchester, UK

[2] Centre for Health Informatics, The University of Manchester, Manchester, UK

[3] Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands

[4] NIHR Manchester Biomedical Research Centre, Manchester University Hospital NHS Foundation Trust, Manchester, UK

[5] Department Pediatric Rheumatology, University Medical Center Utrecht, EuropeanReferenceNetwork-RITA, Netherlands

[6] Centre for Genetics and Genomics Versus Arthritis, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Sciences Centre, Manchester, UK

***Corresponding author**

Dr Stephanie Shoop-Worrall
Postal address: 2.908 Stopford Building, The University of Manchester, Manchester, UK, M13 9PT
Email: Stephanie.shoop-worrall@manchester.ac.uk
Tel: +44 1612757757
Orcid ID: 0000-0002-9441-5535
Word count: 3782

Key words: Item Response Theory, Juvenile Idiopathic Arthritis, Adolescent Rheumatology, Functional Ability

SOURCES OF FUNDING

Funded by the Medical Research Council (UK Grant number: MR/K501311/1) and supported by the NIHR Manchester Biomedical Research Centre and the Versus Arthritis Centres for Excellence in Epidemiology and Genetics/Genomics (UK grant numbers 20380, 20542). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

COMPETING INTERESTS

The authors declare no conflicts of interest.

ABSTRACT

Objectives:

As young people enter adulthood, the interchangeable use of child and adult outcome measures may inaccurately capture changes over time. This study aimed to use item response theory (IRT) to model a continuous score for functional ability that can be used no matter which questionnaire is completed.

Methods:

Adolescents (11-17 yrs) in the UK Childhood Arthritis Prospective Study self-completed an adolescent-Childhood Health Assessment Questionnaire (A-CHAQ) and a Health Assessment Questionnaire (HAQ). Their parents completed the proxy-CHAQ (P-CHAQ). Those with at least two simultaneously completed questionnaires at initial presentation or one year were included.

Psychometric properties of item responses within each questionnaire were tested using Mokken analyses to assess the applicability of IRT modelling. A previously developed IRT model from the Pharmachild-NL registry was validated in CAPS participants. Agreement and correlations between IRT-scaled functional ability scores were tested using intra-class correlations and Wilcoxon signed-ranked tests.

Results:

In 303 adolescents, median age at diagnosis was 13 years and 61% were female. CHAQ scores consistently exceeded HAQ scores.

Mokken analyses demonstrated high scalability, monotonicity and that each questionnaire yielded reliable scores. There was little difference in item response characteristics between adolescents enrolled to CAPS and Pharmachild (maximum item residual 0.08). Significant differences were no longer evident between IRT-scaled HAQ and CHAQ scores.

Conclusion:

IRT modelling allows the direct comparison of function scores regardless of different questionnaires being completed by different people over time. This facilitates ongoing assessment of function as adolescents transfer from paediatric clinics to adult services.

SIGNIFICANCE AND INNOVATIONS

1. Functional ability is a key outcome for adolescents transitioning into adulthood

2. Using Item Response Theory, a common scale for functional ability has been developed and validated

3. Direct comparison of functional ability through adolescence is now feasible using this common scale

INTRODUCTION

Functional ability is an important patient-reported outcome in people with juvenile idiopathic arthritis (JIA), both in childhood and later life [1]. As a young person with JIA moves through adolescence and into adulthood, their functional ability may be assessed using one of three versions of the Health Assessment Questionnaire (HAQ), depending on their age and local practice: the proxy-completed Childhood Health Assessment Questionnaire (P-CHAQ) [2], a self-completed adolescent-CHAQ (A-CHAQ) with the same items as the P-CHAQ but developmentally appropriate re-wording [3] or the self-completed Stanford HAQ, which has fewer items and was originally designed for adults with rheumatoid arthritis [4]. The P-CHAQ was adapted from the HAQ and thus assesses similar domains of functional ability, with additional items for tasks more relevant to young people e.g. writing with a pen/pencil. In addition, a modified HAQ (MHAQ) was developed from the HAQ in order to reduce time burden for both

patients and healthcare professionals. The MHAQ includes one question from each HAQ domain and can be completed in under five minutes by adults with rheumatoid arthritis [5].

Directly comparing scores on these four 'similar' outcome measures is challenging since each questionnaire has unique questions, or items. This may lead to differences in scores that are unrelated to actual differences in underlying functional ability [6]. In addition, they may have been completed by different people over time (e.g. adolescent, parent/carer). Finally, questionnaires may contain missing values, especially when paper and pencil forms are used. These limitations hinder the continuous assessment of functional ability as adolescents mature and are transferred from paediatric to adult care, with previous work demonstrating that these existing questionnaires produce scores that are similar, but not interchangeable, when completed by/for the same young person at the same time point [6].

To continuously assess functional ability over time, a common function scale is needed. It would be difficult to use a single questionnaire for people with JIA of all ages, since some functional tasks are age-specific and different people (carer versus young person) may need to complete the questionnaire over time. One established method used to link scores from different questionnaires to a common scale is item response theory (IRT) [7,8]. Within IRT, item and person characteristics are mapped on the same underlying measurement continuum. These characteristics encompass i) the trait level of the person completing the item (i.e. the level of their functional ability) and ii) the characteristics of the items themselves (e.g. the general difficulty of opening a jar versus a car door). One useful benefit of modelling item responses this way is that it allows the scores to be corrected for these item characteristics [9,10]. This way, a single score can be reflective of underlying functional ability, no matter what questionnaires or items have been completed.

The applications of IRT models are increasingly popular in outcome assessments across various medical fields. For example, in the Patient Reported Outcomes Measurement Information System (PROMIS [11]) project, various item banks have been developed, from which tailored questionnaires with different items and lengths can be derived, with optimal relevance to specific patients [12]. In the patient-reported outcome Rosetta Stone (PROsetta Stone) project, IRT was one method used to link 'legacy measures', those already developed and historically used, with newer PROMIS measures, to allow the retrofitting of existing scores to the newer measures and vice versa [13]. In addition, IRT has previously been used to model latent functional ability across

multiple questionnaires in adults with rheumatoid arthritis [14]. However, to date, its application in JIA, in addition to similar questionnaires that have been sequentially developed, is limited.

Recently, an IRT-based standardised functional ability reporting metric was developed in 16,386 people with inflammatory rheumatic diseases recruited to international registries, including the Pharmachild registry of children and young people with JIA [14]. The standardised functional ability scale developed includes ten commonly used functional ability questionnaires (and their items), including the HAQ, MHAQ and the P-CHAQ and can be used to obtain comparable scores from each of the included questionnaires. It could therefore be used in young people with JIA to obtain comparable physical function scores regardless of the particular functional ability questionnaire used.

The aim of the current study was to examine (1) the applicability of this metric in JIA. This could be assessed by examining the assumptions and fit of the IRT model underlying the common metric in data obtained in a population of adolescents with JIA in the UK; (2) the agreement between IRT-scaled scores obtained using P-CHAQ, A-CHAQ and HAQ in adolescents with JIA and (3) the measurement properties of these questionnaires in this population using non-parametric IRT analyses.

METHODS

Development study population

The Pharmachild-NL registry is a web-based register extracting demographic and clinical data from medical records twice yearly for juvenile arthritis in Utrecht, NL. The cohort has been previously described [14]. This cohort included 1194 prevalent cases of juvenile arthritis who were prescribed methotrexate or biologic therapies and were selected for development of the IRT model. Item responses from the P-CHAQ, HAQ and MHAQ were extracted from CYP contributing these data between 2010 and 2017.

Validation study population

Data were obtained from adolescents enrolled in the Childhood Arthritis Prospective Study (CAPS). CAPS is a longitudinal, UK, multicentre inception cohort following children and young people with inflammatory arthritis with onset before their 16th birthday. Specific inclusion and exclusion criteria for CAPS have been described previously [15]. CAPS has been approved by the

Northwest Multicentre Research Ethics Committee (REC/02/8/104, IRAS 184042) and written informed consent was provided by proxies for all participants; where possible, patient assent was also obtained.

Between January 2004 and January 2015, adolescents aged between 11 and 17 years enrolled in CAPS, were asked to self-complete the A-CHAQ and HAQ and for their proxies to complete the P-CHAQ at the same clinic visit. Only those adolescents with data from at least two of these three questionnaires completed at either initial presentation to paediatric rheumatology (CAPS baseline) or at one year following initial presentation (CAPS 1-year follow-up) were included in the current analysis. MHAQ scores were calculated using existing HAQ scores where available, with one item from each domain included [16].

Additional data collected at baseline from the CAPS cohort included demographic (ethnicity, gender, dates of birth, disease onset and initial presentation) and disease-related variables collected at both baseline and one year (disease category, active joint count, limited joint count, ESR (mm/hr), physician's global assessment of disease (10cm visual analogue scale), proxy global assessment of wellbeing (10cm visual analogue scale)).

Statistical analysis

Calculating CHAQ/HAQ Scores in CAPS data

Item-specific, domain-specific and overall CHAQ/ HAQ scores were calculated using CAPS data at baseline and one year. Due to translation discordance between the UK and Netherlands (NL) CHAQ versions, the UK item regarding running errands (NL 'run a race') was omitted. IRT models are robust to missing item data and overall scores can be compared using a total of the remaining items [14]. To gain an overall score for each questionnaire the largest possible item scores (0-3) within each domain (8 in total) were summed for a possible range of 0 to 24. Dividing by 8 yields a final score ranging from 0 to 3 (increasing scores denote worsening disability). In cases of incomplete data, a final score can be calculated if at least six of eight domains have values, through dividing by the number of domains with available data instead. In this study, use of aids and devices were not considered when calculating domain-specific scores in order to assess the effects of item differences on overall scores.

Assessing IRT assumptions in CAPS data

The IRT model that was used for calibrating the items from each questionnaire to a common function ability scale, the generalised partial credit model [17], has two assumptions: i) unidimensionality: that all items from each functional ability questionnaire relate to the common underlying continuous function variable and ii) monotonicity: that the expected item score functions are monotonically increasing over this latent variable (i.e. the common functional ability scale increases each time an item score increases). Both assumptions were tested via checking goodness of fit of Mokken's model of monotone homogeneity [18]. This is a non-parametric IRT model used to verify that patients can be ordered along an underlying latent variable. The model relies on the same assumptions as the generalised partial credit model. In the Mokken approach, the unidimensionality assumption can be checked using item (H_i) and scale (H) level scalability coefficients. Higher values indicate better scalability. $H > 0.30$ supports unidimensionality and $H > 0.50$ suggests a strong scale [19]. The monotonicity assumption was checked using the *check.monotonicity* function of the Mokken R package. Subsequently, we examined the reliability of the overall scores for each questionnaire using the Molenaar Sijtsma coefficient.

Fitting the IRT model in CAPS data

Differences in item response behaviour between adolescents enrolled in Pharmachild (P-CHAQ) and CAPS (P-CHAQ, A-CHAQ, HAQ) were then examined to assess if the existing item parameters were generalisable. This was completed by testing for differential item functioning (DIF). DIF occurs if adolescents with the same level of functional ability across cohorts have different IRT expected item scores. DIF was examined using Lagrange multiplier statistics and associated effect size statistics [20].

Subsequently, we fitted the previously estimated IRT model in the CAPS data. We tested the fit of the models by calculating the differences between the observed item scores in CAPS and the IRT model predicted scores (i.e. the absolute residuals). Item fit was considered acceptable if an item's score residual was $<\pm 0.2$.

A test characteristic curve and conversion tables were constructed to demonstrate how raw CHAQ, HAQ and MHAQ scores (as scored in this paper with the 19 item HAQ and without the use of aids) can be compared with standardised functional ability scores and/or translated amongst each other. The conversion tables were constructed according to Thissen et al's expected a posteriori (EAP) approach for summed scores, using the Lord Wingerky algorithm [21]. These stand only

where no missing data are evident. To gain more accurate comparisons to latent scores, the converter tool at http://tihealthcare.nl/en/expertise/common-metrics can be used and an app is currently under development.

Evaluating congruence of IRT scores obtained from different functional ability questionnaires

Finally, the comparability of functional ability scores was assessed between IRT-scaled and raw CHAQ and HAQ scores. Pairwise agreement between EAP IRT scores from the four functional ability measures was assessed [22]. The EAP score estimation procedure was chosen because of the sizable flooring effect of the CHAQ/HAQ. Pairwise agreements between overall raw scores and between EAP-modelled IRT scores at baseline were assessed using Bland-Altman plots and compared using Wilcoxon Signed-ranked tests. All analyses were undertaken in Stata 14 (Stata Corp, College Station, TX, USA) and R 3.4.1 (R Core Team, Vienna, Austria).

RESULTS

Patient cohort

A total of 303 adolescents in CAPS had completed at least two of the three full questionnaires at either the baseline (n=178) or one year (n=231) visit. Compared with those adolescents with fewer than two questionnaire responses at either time point (n=77), those included in the study had marginally higher physician global scores (2.5cm vs 3.1cm, p=0.032). There were no differences in age, gender, ethnicity, disease duration, ILAR category, pain or any of the JIA core outcome variables except physician's global scores at baseline between these included and excluded from the study. Available CHAQ/HAQ scores were equivalent between the two groups.

The majority were female (59%) and of white ethnicity (91%). The median age at initial presentation to paediatric rheumatology was 13 years (IQR 12 to 14) with median seven months symptom duration to this point (IQR 4 to 17). The most common disease category was oligoarticular JIA (40%). At this time, adolescents had a median of two active joints and physician and proxy global scores at approximately 3cm on a 10cm visual analogue scale (Table 1).

At baseline, median CHAQ scores were consistent across proxies and adolescents at both baseline (both CHAQ median: 0.6, both IQRs 0.1 to 1.3) and one year (both CHAQ medians: 0.3, both IQRs 0.0, 0.8). HAQ and MHAQ scores consistently scored below those of the CHAQ (baseline

HAQ: 0.5 (IQR 0.0, 1.3), one year HAQ: 0.1 (IQR 0.0, 0.8), baseline MHAQ 0.1 (0.0, 0.5), one year MHAQ 0.0 (0.0, 0.1)) (Table 1).

The CAPS cohort was similar in gender, ethnicity and ILAR distributions to the development population from Pharmachild-NL (65% female, 96% white ethnicity, 48% oligoarthritis). Although Pharmachild-NL included prevalent cases, their age at CHAQ/HAQ completion was comparable (mean 13 years, SD 7 years). Similar to the CAPS cohort, CHAQ scores (median 0.5, IQR 0.1, 1.0) were higher than HAQ (median 0.4, IQR 0.0, 0.9) and MHAQ scores (median 0.1, IQR 0.0, 0.5).

Checking IRT assumptions and the psychometric properties of CHAQ/HAQ scores in CAPS

The IRT model assumptions held for each functional ability measure, suggesting that an IRT approach was applicable to functional ability in JIA using these questionnaires. Strong scalability and unidimensionality were evident for overall P-CHAQ, A-CHAQ and HAQ scores at both baseline and one year (all H>0.5, all SE <0.1). Item specific associations with the latent functional ability variable varied between items within questionnaires in terms of both scalability coefficients (H_i ranges: P-CHAQ 0.3 to 0.7, A-CHAQ 0.3 to 0.7, HAQ 0.4 to 0.7) and concordance coefficients (concordance coefficient ranges: P-CHAQ 0.4 to 0.8, A-CHAQ 0.4 to 0.8, HAQ 0.5 to 0.8). There were no violations to monotonicity and reliability for each questionnaire at each time point was high (all reliability coefficients ≥ 0.95) (Supplementary Table 1).

Assessing differences in item response behaviour between CAPS and Pharmachild and IRT model fit

The DIF analyses are summarized in Supplementary Table 2 and suggested no great differences in how adolescents in CAPS and Pharmachild respond to the items. In general, the observed HAQ, P-CHAQ and A-CHAQ average item scores were similar to the average item scores predicted by a joint IRT calibration of the CAPS and Pharmachild data, with all residuals < 0.10, and only 1% of item residuals exceeding ±0.05 (Supplementary Table 2).

Subsequently, the fit of the item parameters calibrated in Oude Voshaar et al [14] were evaluated in CAPS data. Again, the model-predicted average item scores were generally close to the average item scores observed in the CAPS data, with residuals consistently falling below 0.2 across all questionnaires (Supplementary Table 2).

Directly comparing latent functional ability across different questionnaires with different completers

Figure 1 shows how the CHAQ and HAQ scores relate to the standardised physical function score metric. In addition, Supplementary Table 3 allows the direct comparison of CHAQ, HAQ and MHAQ scores to this score metric. Increasing values on the standardised function scores indicate better functional ability. The figure and conversion tables can be used to compare CHAQ scores to the standardised physical function scores and re-translate to HAQ scores if needed. However, this exact relationship only applies where no missing values are evident.

Agreement between scores across modelling techniques

Bland-Altman plots demonstrated greater agreement between IRT-scaled than raw scores, demonstrated by narrower limits of agreement and greater centrality around a mean difference of zero for all pairs of scores (Supplementary Figure 1). The majority of pairings had significant differences between raw scores and non-significant differences between IRT-scaled scores. In addition, T-values were lower for all IRT-scaled pairings than raw scores, with the exception of the P-CHAQ vs A-CHAQ at baseline (Table 2).

DISCUSSION

Upon reaching adolescence and following transfer from paediatric to adult care, outcomes in adolescents with JIA are measured using self-completed questionnaires rather than via proxy reports. For functional ability, this often means the HAQ is used instead of the P-CHAQ, with the potential intermediate use of the A-CHAQ. Previous work has demonstrated high correlation but only moderate agreement between raw scores using these three measures [6,23,24]. Therefore, assuming that the scores are interchangeable may result in the false assumption of an improvement in ability where no such change had occurred, based only on the choice of questionnaire. Similarly, longitudinal outcome studies in JIA which capture data across adolescence and young adulthood [25] may also make incorrect conclusions about functional ability over this period if the choice of measure is not considered. The current study demonstrated the applicability of IRT modelling using CHAQ/HAQ item responses. This could be used to understand functional ability in young people with JIA over longer periods of time, retrospectively scale functional ability

scores from completed studies in order to increase standardised comparison and allow for the interpretation of incomplete functional ability questionnaires. Models initially developed in an international cohort including children and young people with JIA were validated in a UK multicentre inception cohort. This resulted in greater agreement between overall IRT-scaled scores than between raw scores. The IRT models presented therefore allow the direct comparison of P-CHAQ, A-CHAQ, HAQ and/or MHAQ scores over time, with an underlying latent variable score and with each other. Further research using any of these measures in JIA should report scaled values alongside raw scores, to allow direct comparison of functional ability between cohorts that may have used different questionnaires.

The psychometric properties of CHAQ/HAQ/MHAQ scores in relation to IRT modelling have rarely been assessed. Previous smaller studies including prevalent cases of JIA have found it hard to estimate stable item parameters [26,27]. In both studies, small sample sizes, in addition to the prevalent flooring effect of the questionnaires, limited the accuracy of generated parametric-IRT (Rasch) parameters. One study resorted to combining the 'with much difficulty' and 'unable to do' CHAQ categories in order to force Rasch model fit [26]. To overcome these issues, the current study employed non-parametric IRT models in a population at least twice the sample size than in previous works. These models do not rely on estimated parameters to study the measurement properties of the included scales. Our results therefore provide useful additional information about the psychometric properties of the evaluated questionnaires. We were able to show that all items on the P-CHAQ, A-CHAQ and HAQ relate to a single underlying functional ability variable and that each instrument yields highly reliable scores.

Once the applicability of IRT modelling to each of the three questionnaires had been confirmed, the current study was able to validate existing IRT models developed in young people and adults with JIA in the Pharmachild-NL registry. Previously fitted models successfully summarised the item responses given by adolescents in CAPS. Thus, the results should generalise across other cohorts of patients with JIA, regardless of which questionnaire has been completed. The utility of the models was demonstrated in the increased agreement between pairs of overall scores under these models compared to raw scores, with the former adjusting for item characteristics.

If complete data are available, the conversion table (Supplementary Table 3) and figure (Figure 1) can be used to access latent functional ability scores. In cases of missing data, or to convert entire datasets at once, the now externally-validated models are available at

http://tihealthcare.nl/en/expertise/common-metric and can be used to directly access latent functional ability scores for individual patients or cohorts of patients for both clinical and research purposes.

Limitations to the study include the small differences between CHAQ and HAQ items, few of which were entirely unique to each questionnaire. Despite the differences between questionnaire scores being greater than the minimal clinically important differences in functional ability [28,29], this analysis did not demonstrate the full possibilities of IRT modelling. Further applications include its ability to model other functional ability questionnaires with unique items, such as CHAQ compared with the functional ability questions within the Juvenile Arthritis Multidimensional Assessment Report (JAMAR) [30]. With increasing differences in questionnaires measuring the same disease construct, greater differences between raw scores and IRT-scaled scores would be evident. However, for this study, CHAQ and HAQ scores have been assumed interchangeable and even with these small changes between questionnaire items, the current study was able to demonstrate i) greater agreement between IRT-scaled compared with raw scores, ii) scores that are not biased in the presence of incomplete answers compared with raw scores and iii) the ability to directly compare scores from any of the questionnaires with an underlying construct variable. In clinical practice, these models facilitate direct comparison of CHAQ scores with HAQ scores upon switching of questionnaires during adolescence. This includes the MHAQ, with lesser burden on adolescents since only six items on the HAQ are required for a total score, taking fewer than five minutes to complete [5], with young people previously reporting that the CHAQ was burdensome in length [31]. Beyond this, functional ability questionnaires can be tailored to each young person based on personalised relevance from a functional ability item bank such as PROMIS [11]. IRT modelling would then allow for the direct comparison of functional ability over time, even when different items have been completed from these different questionnaires.

Further limitations include that functional ability of the tested cohort was, on average, low to moderate and thus few very high CHAQ/HAQ scores contributed to the models. The flooring effect of these questionnaires is well known [2], with upper quartile scores extended to only 1.3 out of 3.0 even at initial presentation to paediatric rheumatology. Whilst few patients experienced severe limitations in functional ability, this validation cohort represents a generalisable sample of adolescents with newly-diagnosed JIA, including those across all ILAR categories. Finally, the

current study was able to demonstrate a direct comparison between latent functional ability and a proxy-completed P-CHAQ. However, it is often evident that young people with JIA complete the P-CHAQ themselves, particularly where the A-CHAQ and HAQ are not available. No adolescents in this study self-completed the P-CHAQ. However, the lack of differences in item responses between the proxy completed P-CHAQ and adolescent completed A-CHAQ meant that the current study could combine these questionnaires to a single CHAQ score. Thus, the CHAQ model presented should be able to adequately incorporate self-completed P-CHAQ scores. Finally, these data were collected as part of an observational 'real-world' research study. As in any longitudinal observational study, clinical and demographic data are often missing. To allow for adequate validation of the IRT model, we required at least two of the CHAQ/HAQ forms to have been completed. Available CHAQ/HAQ scores were equivalent between adolescents included and excluded from the study.

CONCLUSION

P-CHAQ, A-CHAQ and HAQ scores can be directly compared to latent functional ability using IRT modelling. This will greatly aid the direct comparison of functional ability across the JIA disease course when adolescents are transferred from paediatric to adult rheumatology services. In additional, scores from different study populations using different functional ability questionnaires can be directly compared, and longer-scale studies can now feasibly compare functional ability even if questionnaires have missing items and/or adolescents switch questionnaires throughout the study.

ACKNOWLEDGEMENTS

The authors thank all of the children, young people and their guardians involved in CAPS and Pharmachild in addition to all clinical staff and administrators. We also thank the data management team at the University of Manchester, UK.

REFERENCES

(1) Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum. 1997;40:1202-9.

(2) Nugent J, Ruperto N, Grainger J, Machado C, Sawhney S, Baildam E, et al. The British version of the Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ). Clin Exp Rheumatol. 2001;19:S163-S167.

(3) Shaw KL, Southwood TR, McDonagh JE. Growing up and moving on in rheumatology: parents as proxies of adolescents with juvenile idiopathic arthritis. Arthritis Rheum. 2006;55:189-98.

(4) Kirwan JR, Reeback JS. Stanford Health Assessment Questionnaire modified to assess disability in British patients with rheumatoid arthritis. Br J Rheumatol. 1986;25:206-9.

(5) Maska L, Anderson J, Michaud K. Measures of functional status and quality of life in rheumatoid arthritis: Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment Questionnaire II (HAQ-II), Improved Health Assessment Questionnaire (Improved HAQ), and Rheumatoid Arthritis Quality of Life (RAQoL). Arthritis Care Res (Hoboken). 2011;63 Suppl 11:S4-13.

(6) Shoop-Worrall SJW, Hyrich KL, Verstappen SM, Sergeant JC, Baildam E, Chieng A, et al. Comparing Proxy, Adolescent and Adult Assessments of Functional Ability in Adolescents with Juvenile Idiopathic Arthritis. Arthritis Care Res (Hoboken). 2019.

(7) Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol. 2009;5:27-48.

(8) Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38:II28-II42.

(9) van der Linden WJ, Glas CAW. Elements of Adaptive Testing. Springer-Verlag New York; 2010.

(10) Stocking ML, Frederic LM. Developing a Common Metric in Item Response Theory. Applied Psychological Measurement. 1983;7:201-10.

(11) 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;63:1179-94.

(12) Oude Voshaar MA, ten Klooster PM, Glas CA, Vonkeman HE, Taal E, Krishnan E, et al. Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. PLos One. 2014;9:e92367.

(13) Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess. 2014;26:513-27.

(14) 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. 2018.

(15) Adib N, Hyrich K, Thornton J, Lunt M, Davidson J, Gardner-Medwin J, et al. Association between duration of symptoms and severity of disease at first presentation to paediatric rheumatology: results from the Childhood Arthritis Prospective Study. Rheumatology (Oxford). 2008;47:991-5.

(16) Blalock SJ, Sauter SVH, Devellis RF. The modified health assessment questionnaire difficulty scale. A health status measure revisited. Arthritis Rheumatol. 1990;3:182-8.

(17) Muraki E. A Generalized Partial Credit Model. In: ven der Linden WJ, Hambleton RK, editors. Handbook of Modern Item Response Theory.New York: Springer; 1997. p. 153-64.

(18) Molenaar IW. Nonparametric Models for Polytomous Responses. In: van der Linden WJ, Hambleton RK, editors. Handbook of Modern Item Response Theory.New York: Springer; 1997. p. 369-80.

(19) Sijtsma K, van der Ark LA. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. Br J Math Stat Psychol. 2017;70:137-58.

(20) Glas CAW. Detection of Differential item Functioning Using Lagrange Multiplier Tests. Statistica Sinica. 1998;8:647-67.

(21) Thissen D, Pommerich M, Billeaud K, Williams VSL. Item response theory for scores on tests including polytomous items with ordered responses. Applied Psychological Measurement. 1995;19:39-49.

(22) Warm TA. Weighted likelihood estimation of ability in item response theory. Psychometrika. 1989;54:427-50.

(23) Van Pelt PA, Kruize AA, Goren SS, Van Der Net J, Uiterwaal C, Kuis W, et al. Transition of rheumatologic care, from teenager to adult: which health assessment questionnaire can be best used? Clin Exp Rheumatol. 2010;28:281-6.

(24) Lal SD, McDonagh J, Baildam E, Wedderburn LR, Gardner-Medwin J, Foster HE, et al. Agreement between proxy and adolescent assessment of disability, pain, and well-being in juvenile idiopathic arthritis. J Pediatr. 2011;158:307-12.

(25) Gore FM, Bloem PJ, Patton GC, Ferguson J, Joseph V, Coffey C, et al. Global burden of disease in young people aged 10–24 years: a systematic analysis. Lancet. 2011;377:18-24.

(26) Pouchot J, Ecosse E, Coste J, Guillemin F. Validity of the childhood health assessment questionnaire is independent of age in juvenile idiopathic arthritis. Arthritis Rheum. 2004;51:519-26.

(27) Tennant A, Kearns S, Turner F, Wyatt S, Haigh R, Chamberlain MA. Measuring the function of children with juvenile arthritis. Rheumatology (Oxford). 2001;40:1274-8.

(28) Brunner HI, Klein-Gitelman MS, Miller MJ, Barron A, Baldwin N, Trombley M, et al. Minimal clinically important differences of the childhood health assessment questionnaire. J Rheumatol. 2005;32:150-61.

(29) Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. J Rheumatol. 2009;36:254-9.

(30) Filocamo G, Consolaro A, Schiappapietra B, Dalpra S, Lattanzi B, Magni-Manzoni S, et al. A new approach to clinical care of juvenile idiopathic arthritis: the Juvenile Arthritis Multidimensional Assessment Report. J Rheumatol. 2011;38:938-53.

(31) Parsons S, Thomson W, Cresswell K, Starling B, McDonagh JE, Barbara Ansell National Network for Adolescent Rheumatology (BANNAR). What do young people with rheumatic conditions in the UK think about research involvement? A qualitative study. Pediatr Rheumatol Online J. 2018;16:35.

TABLE AND FIGURE LEGENDS

Table 1. Baseline characteristics of the cohort

Table 2. Significant differences between pairwise functional ability questionnaires

Figure 1. A test characteristic curve demonstrating how latent functional ability can be modelled using either/all of CHAQ, HAQ and MHAQ scores.

TABLES

Table 1. Baseline characteristics of the cohort

Characteristic	Percent complete data (%)	N(%) or median (IQR)
Female	100	180 (59)
White or Caucasian	97	267 (91)
Age at onset (years)	97	12 (11, 13)
Age at first presentation (years)	100	13 (12, 14)
Symptom duration at first paediatric rheumatology appointment (months)	98	7 (4, 17)
ILAR category:	100	
Systemic		20 (7)
Oligoarticular		120 (40)
RF- Polyarticular		56 (18)
RF+ Polyarticular		20 (7)
Enthesitis-related		30 (10)
Psoriatic		30 (10)
Undifferentiated		27 (9)
Core outcome variables at baseline:		
Active joint count	90	2 (1, 6)
Limited joint count	90	1 (1, 4)
ESR (mm/hr)	70	17 (6, 54)
Physician's global assessment (cm)	64	3.1 (1.7, 5.4)
Proxy global assessment of wellbeing (cm)	77	2.7 (0.7, 5.1)
Functional ability at baseline*:		
P-CHAQ	87	0.625 (0.125, 1.250)
A-CHAQ	89	0.625 (0.125, 1.250)

Characteristic	Percent complete data (%)	N(%) or median (IQR)
HAQ	87	0.500 (0.000, 1.250)
MHAQ	87	0.125 (0.000, 0.500)
Functional ability at one year*:		
P-CHAQ	90	0.250 (0.000, 0.750)
A-CHAQ	89	0.250 (0.000, 0.750)
HAQ	93	0.125 (0.000, 0.750)
MHAQ	93	0.000 (0.000, 0.125)

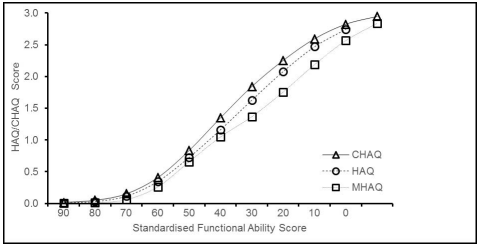
*Of those that had ≥2 complete functional ability questionnaires at time point. IQR: Interquartile range, ILAR: International League of Associations for Rheumatology, RF: Rheumatoid factor, ESR: Erythrocyte sedimentation rate, P-CHAQ: Proxy Childhood Health Assessment Questionnaire, A-CHAQ: Adolescent-CHAQ, HAQ: Health Assessment Questionnaire, MHAQ: modified HAQ.

Table 2. Significant differences between pairwise functional ability questionnaires

Questionnaire comparison	Model	Baseline				One year			
		n	% ceiling*[1]	T*[2]	P-value*[2]	n	% ceiling*[1]	T*[2]	P-value*[2]
PCHAQ vs ACHAQ	Raw data	136	19.3	1.3	0.196	183	41.3	0.6	0.580
	IRT: EAP	136		1.5	0.138	183		0.2	0.843
PCHAQ vs HAQ	Raw data	133	25.7	3.2	0.002	192	45.3	1.3	0.205
	IRT: EAP	133		1.6	0.109	192		-0.2	0.851

PCHAQ vs MHAQ	Raw data	133	23.1	8.7	<0.001	192	43.3	7.1	<0.001
	IRT: EAP	133		1.9	0.059	192		0.8	0.425
ACHAQ vs HAQ	Raw data	136	32.1	3.2	0.002	191	51.2	1.1	0.263
	IRT: EAP	136		2.1	0.036	191		0.0	0.978
A-CHAQ vs MHAQ	Raw data	136	46.4	10.1	<0.001	191	61.7	7.1	<0.001
	IRT: EAP	136		2.6	0.012	191		0.8	0.432
HAQ vs MHAQ	Raw data	156	24.3	9.9	<0.001	218	42.8	9.1	<0.001
	IRT: EAP	156		1.0	0.340	218		2.0	0.052

*1: % 0 on both scores, *2: Wilcoxon signed-rank test. P-CHAQ: Proxy Childhood Health Assessment Questionnaire, A-CHAQ: Adolescent-CHAQ, HAQ: Health Assessment Questionnaire, MHAQ: modified HAQ, IRT: Item Response Theory, EAP: Expected a priori


3.0
2.5
2.0
1.5
1.0
0.5
0.0
HAQ/CHAQ Score
90
80
70
60
50
40
30
20
10
0
Standardised Functional Ability Score
CHAQ
HAQ
MHAQ