Arthritis Care & Research
Vol. 0, No. 0, Month 2021, pp 1–8
DOI 10.1002/acr.24204

American College of Rheumatology
Empowering Rheumatology Professionals

Common Functional Ability Score for Young People With Juvenile Idiopathic Arthritis

Stephanie J. W. Shoop-Worrall,[1] Martijn A. H. Oude Voshaar,[2] Janet E. McDonagh,[3] Mart A. F. J. Van de Laar,[2] Nico Wulffraat,[4] Wendy Thomson,[3] Kimme L. Hyrich,[3] and Suzanne M. M. Verstappen[3]

Objective. 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 (ages 11–17 years) in the UK Childhood Arthritis Prospective Study (CAPS) self-completed an adolescent Childhood Health Assessment Questionnaire (CHAQ) and a Health Assessment Questionnaire (HAQ). Their parents answered the proxy-completed CHAQ. Those children with at least 2 simultaneously completed questionnaires at initial presentation or 1 year were included. Psychometric properties of item responses within each questionnaire were tested using Mokken analyses to assess the applicability of IRT modeling. A previously developed IRT model from the Pharmachild-NL registry from The Netherlands was validated in CAPS participants. Agreement and correlations between IRT-scaled functional ability scores were tested using intraclass correlations and Wilcoxon's signed rank tests.

Results. In 303 adolescents, the median age at diagnosis was 13 years, and 61% were female. CHAQ scores consistently exceeded HAQ scores. Mokken analyses demonstrated high scalability, monotonicity, and the fact that each questionnaire yielded reliable scores. There was little difference in item response characteristics between adolescents enrolled in CAPS and Pharmachild-NL (maximum item residual 0.08). Significant differences were no longer evident between IRT-scaled HAQ and CHAQ scores.

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

INTRODUCTION

Functional ability is an important patient-reported outcome in individuals 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 1 of 3 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 rewording (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 to reduce the time burden for both patients and health care professionals. The MHAQ includes 1 question from each HAQ domain and can be completed in under 5 minutes by adults with rheumatoid arthritis (5).

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Supported by the Medical Research Council (UK grant MR/K501311/1) and by the NIHR Manchester Biomedical Research Centre and the Versus Arthritis Centres for Excellence in Epidemiology and Genetics/Genomics (UK grants 20380 and 20542).

[1]Stephanie J. W. Shoop-Worrall, PhD: University of Manchester and Manchester Academic Health Sciences Centre, Manchester, UK; [2]Martijn A. H. Oude Voshaar, PhD, Mart A. F. J. Van de Laar, MD, PhD: University of Twente, Enschede, The Netherlands; [3]Janet E. McDonagh, MD, FRCP, Wendy Thomson, PhD, Kimme L. Hyrich, MD, PhD, FRCPC, Suzanne M. M. Verstappen, PhD: University of Manchester and Manchester Academic Health Sciences Centre, NIHR Manchester Biomedical Research Centre, and Manchester University Hospital NHS Foundation Trust, Manchester, UK; [4]Nico Wulffraat, MD, PhD: University Medical Center Utrecht and European Reference Network–RITA, Utrecht, The Netherlands.

No potential conflicts of interest relevant to this article were reported.

Address correspondence to Stephanie J. W. Shoop-Worrall, PhD, 2.908 Stopford Building, The University of Manchester, Manchester, UK, M13 9PT. Email: Stephanie.shoop-worrall@manchester.ac.uk.

Submitted for publication December 13, 2019; accepted in revised form March 31, 2020.

SIGNIFICANCE & INNOVATIONS

- Functional ability is a key outcome for adolescents transitioning into adulthood.
- Using item response theory, a common scale for functional ability has been developed and validated.
- Direct comparison of functional ability through adolescence is now feasible using this common scale.

Directly comparing scores on these 4 similar outcome measures is challenging, since each questionnaire has unique questions, or items. This diversity may lead to differences in scores that are unrelated to actual differences in underlying functional ability (6). In addition, the questionnaires may have been completed by different people over time (e.g., adolescent, parent/care giver). 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 pediatric to adult care, with previous research 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. Using a single questionnaire for individuals with JIA of all ages would be difficult, since some functional tasks are age-specific and different people (care giver 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 the trait level of the person completing the item (i.e., the level of their functional ability), and the characteristics of the items themselves (e.g., the general difficulty of opening a jar versus a car door). One useful benefit of modeling item responses this way is that the modeling 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 standardized functional ability reporting metric was developed in 16,386 people with inflammatory rheumatic diseases recruited to international registries, including the Pharmachild-NL registry of children and young people with JIA (14). The standardized functional ability scale developed includes 10 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, which 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 nonparametric IRT analyses.

SUBJECTS AND 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, The Netherlands. The cohort has been previously described (14). This cohort included 1,194 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 young people 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, multicenter 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 ages 11–17 years who were 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 2 of these 3 questionnaires completed at either initial presentation to pediatric rheumatology (CAPS baseline) or at 1 year following the initial presentation (CAPS 1-year follow-up)

were included in the current analysis. MHAQ scores were calculated using existing HAQ scores where available, with 1 item from each domain included (16).

Additional data collected at baseline from the CAPS cohort included demographic (ethnicity, sex, date of birth, disease onset, and initial presentation) and disease-related variables collected at both baseline and 1 year (disease category, active joint count, limited joint count, erythrocyte sedimentation rate [mm/hour], physician's global assessment of disease [10-cm visual analog scale], and proxy global assessment of well-being [10-cm visual analog 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 1 year. Due to translation discordance between the UK and The Netherlands CHAQ versions, the UK item regarding running errands (Netherlands: 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 0–24. 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 6 of 8 domains have values, through dividing by the number of domains with available data instead. In this study, the use of aids and devices was 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 generalized partial credit model (17), has 2 assumptions: 1) unidimensionality: that all items from each functional ability questionnaire relate to the common underlying continuous function variable; and 2) 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 by checking the goodness-of-fit of Mokken's model of monotone homogeneity (18). This is a nonparametric IRT model used to verify that patients can be ordered along an underlying latent variable. The model relies on the same assumptions as the generalized partial-credit model. In the Mokken approach, the unidimensionality assumption can be checked using item-level (H_i) and scale-level (H) 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 behavior between adolescents enrolled in Pharmachild-NL (P-CHAQ) and CAPS (P-CHAQ, A-CHAQ, HAQ) were then examined to assess whether the existing item parameters were generalizable. 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 less than ±0.2.

A test characteristic curve and conversion tables were constructed to demonstrate how raw CHAQ, HAQ, and MHAQ scores (as scored in this article with the 19-item HAQ and without the use of aids) can be compared with standardized functional ability scores and/or translated among each other. The conversion tables were constructed according to the expected a posteriori (EAP) approach of Thissen et al 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 4 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-modeled IRT scores at baseline were assessed using Bland-Altman plots and compared using Wilcoxon's signed rank tests. All analyses were undertaken in Stata software version 14, and R version 3.4.1.

RESULTS

Patient cohort. A total of 303 adolescents in CAPS had completed at least 2 of the 3 full questionnaires at either the baseline (n = 178) or 1 year visit (n = 231). Compared with those adolescents with fewer than 2 questionnaire responses at either time point (n = 77), those included in the study had marginally higher physician global scores (2.5 cm versus 3.1 cm; $P = 0.032$). There were no differences in age, sex, ethnicity, disease duration, International League of Associations for Rheumatology (ILAR) category, pain, or any of the JIA core outcome variables except physician's global scores at baseline between those included

and excluded from the study. Available CHAQ/HAQ scores were equivalent between the 2 groups.

The majority of study participants were female (59%) and of white ethnicity (91%). The median age at initial presentation to pediatric rheumatology was 13 years (interquartile range [IQR] 12–14) with median 7 months symptom duration to that point (IQR 4–17). The most common disease category was oligoarticular JIA (40%). At that time, adolescents had a median of 2 active joints and physician and proxy global scores at ~3 cm on a 10-cm visual analog scale (Table 1).

At baseline, median CHAQ scores were consistent across proxies and adolescents at both baseline (both CHAQ medians 0.6, both IQRs 0.1–1.3) and 1 year (both CHAQ medians 0.3, both IQRs 0.0–0.8). HAQ and MHAQ scores consistently ranked below those of the CHAQ: baseline HAQ 0.5 (IQR 0.0–1.3), 1-year HAQ 0.1 (IQR 0.0–0.8), baseline MHAQ 0.1 (IQR 0.0–0.5), 1-year MHAQ 0.0 (IQR 0.0–0.1) (Table 1).

Table 1. Baseline characteristics of the cohort (n = 303)*

Characteristic	Complete data, %	Value
Female sex, no. (%)	100	180 (59)
White or Caucasian, no. (%)	97	267 (91)
Age at onset, years	97	12 (11–13)
Age at first presentation, years	100	13 (12–14)
Symptom duration at first pediatric rheumatology appointment, months	98	7 (4–17)
ILAR category, no. (%)	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/hour	70	17 (6–54)
Physician's global assessment, cm	64	3.1 (1.7–5.4)
Proxy global assessment of well-being, 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)
HAQ	87	0.500 (0.000–1.250)
MHAQ	87	0.125 (0.000–0.500)
Functional ability at 1 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)

* Values are the median (interquartile range) unless indicated otherwise. A-CHAQ = adolescent Childhood Health Assessment Questionnaire; ESR = erythrocyte sedimentation rate; HAQ = Health Assessment Questionnaire; ILAR = International League of Associations for Rheumatology; MHAQ = modified HAQ; P-CHAQ = proxy CHAQ; RF = rheumatoid factor.
† Of those patients who had ≥2 complete functional ability questionnaires at the time point.

The CAPS cohort was similar in sex, ethnicity, and ILAR distributions to the development population from the Pharmachild-NL registry (65% female, 96% white ethnicity, 48% oligoarthritis). Although Pharmachild-NL included prevalent cases, patient age at CHAQ/HAQ completion was comparable (mean ± SD 13 ± 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 1 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 0.3–0.7, A-CHAQ 0.3–0.7, HAQ 0.4–0.7) and concordance coefficients (coefficient ranges: P-CHAQ 0.4–0.8, A-CHAQ 0.4–0.8, HAQ 0.5–0.8). There were no violations to monotonicity, and the reliability for each questionnaire at each time point was high (all reliability coefficients ≥0.95) (see Supplementary Table 1, available on the *Arthritis Care & Research* website at http://onlinelibrary.wiley.com/doi/10.1002/acr.24204/abstract).

Assessing differences in item response behavior between CAPS and Pharmachild-NL and IRT model fit. The DIF analyses are summarized in Supplementary Table 2, available on the *Arthritis Care & Research* website at http://onlinelibrary.wiley.com/doi/10.1002/acr.24204/abstract, and suggested no great differences in how adolescents in CAPS and Pharmachild-NL responded 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-NL data, with all residuals <0.10, and only 1% of item residuals exceeding ±0.05 (see Supplementary Table 2, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.24204/abstract).

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 (see 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 standardized physical function score metric. In addition,

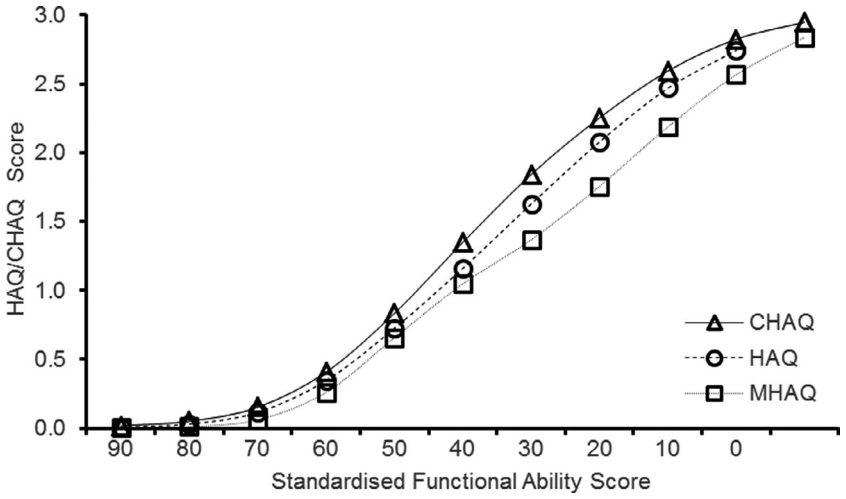

Figure 1. A test characteristic curve demonstrating how latent functional ability can be modeled using either/all of the Childhood Health Assessment Questionnaire (CHAQ), Health Assessment Questionnaire (HAQ), and modified HAQ (MHAQ) scores.

Supplementary Table 3, available on the *Arthritis Care & Research* website at http://onlinelibrary.wiley.com/doi/10.1002/acr.24204/abstract, allows the direct comparison of CHAQ, HAQ, and MHAQ scores to this score metric. Increasing values on the standardized function scores indicate better functional ability. The figure and conversion tables can be used to compare CHAQ scores to the standardized physical function scores and to retranslate to HAQ scores if needed. However, this exact relationship only applies where no missing values are evident.

Agreement between scores across modeling 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 (see Supplementary Figure 1, available on the *Arthritis Care & Research* website at http://onlinelibrary.wiley.com/doi/10.1002/acr.24204/abstract). The majority of pairings had significant differences between raw scores and nonsignificant 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 versus A-CHAQ at baseline (Table 2).

DISCUSSION

Upon reaching adolescence and following transfer from pediatric to adult care, outcomes and 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 3 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 that 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 modeling 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 to increase standardized 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 multicenter 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 modeling have rarely been assessed. Previous smaller studies including prevalent cases of JIA have found estimating stable item parameters to be difficult (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 to force Rasch model fit (26). To overcome these issues, the current

Table 2. Significant differences between pairwise functional ability questionnaires*

	Baseline				One year			
Questionnaire comparison, model	No.	% ceiling†	T‡	*P*‡	No.	% ceiling†	T‡	*P*‡
P-CHAQ vs. A-CHAQ								
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
P-CHAQ 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
P-CHAQ 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
A-CHAQ 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

* A-CHAQ = adolescent Childhood Health Assessment Questionnaire; EAP = expected a priori; HAQ = Health Assessment Questionnaire; IRT = item response theory; MHAQ = modified HAQ; P-CHAQ = proxy CHAQ.
† Percentage 0 on both scores.
‡ Wilcoxon's signed rank test.

study employed nonparametric 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 modeling to each of the 3 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 summarized the item responses given by adolescents in CAPS. Thus, the results should generalize 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 (see Supplementary Table 3, available on the *Arthritis Care & Research* website at http://onlinelibrary.wiley.com/doi/10.1002/acr.24204/abstract) and figure (Figure 1) can be used to access latent functional ability scores. In cases of missing data, or to convert entire data sets 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 minimum clinically important differences in functional ability (28,29), this analysis did not demonstrate the full possibilities of IRT modeling. 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 (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 1) greater agreement between IRT-scaled compared with raw scores, 2) scores that are not biased in the presence of incomplete answers compared with raw scores, and 3) 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 5 items on the HAQ are required for a total score, taking fewer than 5 minutes to complete (5), with young people previously reporting that the CHAQ was burdensome in length (31). Beyond this advantage, functional ability questionnaires can be tailored to each young person based on personalized relevance from a functional ability item bank such as PROMIS (11). IRT modeling 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 the fact 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 of 3.0 even at initial presentation to pediatric rheumatology. While few patients experienced severe limitations in functional ability, this validation cohort represents a generalizable 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 2 of the CHAQ/HAQ forms to have been completed. Available CHAQ/HAQ scores were equivalent between adolescents included and excluded from the study.

P-CHAQ, A-CHAQ, and HAQ scores can be directly compared to latent functional ability using IRT modeling. This will greatly aid the direct comparison of functional ability across the JIA disease course when adolescents are transferred from pediatric 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.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Shoop-Worrall had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Shoop-Worrall, Oude Voshaar, McDonagh, Hyrich, Verstappen.

Acquisition of data. Oude Voshaar, McDonagh, Van de Laar, Wulffraat, Thomson, Hyrich, Verstappen.

Analysis and interpretation of data. Shoop-Worrall, Oude Voshaar, Hyrich, Verstappen.

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:S163–7.
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 SJ, 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) 2020;72:517–24.
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 CA. Elements of adaptive testing. New York: Springer-Verlag; 2010.
10. Stocking ML, Frederic LM. Developing a common metric in item response theory. Appl Psychol Meas 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 MA, 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;28:187–97.
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 SV, 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 CA. Detection of differential item functioning using Lagrange Multiplier Tests. Stat Sin 1998;8:647–67.
21. Thissen D, Pommerich M, Billeaud K, Williams VS. Item response theory for scores on tests including polytomous items with ordered responses. Appl Psychol Meas 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.