Published in final edited form as:
Qual Life Res. 2017 March ; 26(3): 565–586. doi:10.1007/s11136-016-1468-2.

Establishing Clinical Meaning and Defining Important Differences for Patient Reported Outcomes Measurement Information System (PROMIS®) Measures in Juvenile Idiopathic Arthritis Using Standard Setting with Patients, Parents, and Providers

Esi Morgan, MD, MSCE[1,2], **Constance A. Mara, MA, PhD**[3], **Bin Huang, PhD**[1,4], **Kimberly Barnett, BS**[3], **Adam C. Carle, MA, PhD**[1,3,7], **Jennifer E. Farrell, MA**[2], and **Karon F. Cook, PhD**[6]

[1]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH

[2]Division of Rheumatology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[3]James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[4]Division of Behavioral Medicine and Child Psychology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[5]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

[6]Department of Medical Social Sciences, Northwestern University, Chicago, IL

[7]Department of Psychology, University of Cincinnati College of Arts and Sciences, Cincinnati, OH

Abstract

Background—Patient Reported Outcomes Measurement Information System (PROMIS) measures are used increasingly in clinical care. However, for juvenile idiopathic arthritis (JIA), scores lack a framework for interpretation of clinical severity, and minimally important differences (MID) have not been established, which are necessary to evaluate the importance of change.

Methods—We identified clinical severity thresholds for pediatric PROMIS measures of mobility, upper extremity function (UE), fatigue, and pain interference working with adolescents with JIA, parents of JIA patients, and clinicians, using a standard setting methodology modified from educational testing. Item parameters were used to develop clinical vignettes across a range of symptom severity. Vignettes were ordered by severity, and panelists identified adjacent vignettes considered to represent upper and lower boundaries separating category cut points (i.e., from none/mild problems to moderate/severe). To define MIDs, panelists reviewed a full score report for the vignettes and indicated which items would need to change and by how much to represent "just enough improvement to make a difference".

Corresponding Author: Esi Morgan, MD, MSCE, Division of Rheumatology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 4010, Cincinnati, OH 45229. esi.morgan_dewitt@cchmc.org.

Results—For fatigue and UE, cut points among panels were within 0.5 SD of each other. For mobility and pain interference, cut-scores among panels were more divergent, with parents setting the lowest cut-scores for increasing severity. The size of MIDs varied by stakeholders (parents estimated largest, followed by patients, then clinicians). MIDs also varied by severity classification of the symptom.

Conclusions—We estimated clinically relevant severity cut points and MIDs for PROMIS measures for JIA from the perspectives of multiple stakeholders and found notable differences in perspectives.

Keywords

PROMIS; patient-reported outcomes; Item Response Theory (IRT); psychometric methods; juvenile idiopathic arthritis

Juvenile idiopathic arthritis (JIA) is a chronic condition without cure that is associated with impaired physical function, chronic pain, and fatigue [1–4]. Monitoring disease progress, treatment effectiveness, and long term outcomes is served by capturing patient reports of symptoms and disease impact. Patient reported outcomes (PROs) are used increasingly in clinical care [5, 6], but they lack clinical utility without meaningful frameworks to interpret scores. As PROs move from the realm of clinical research and clinical trials to use in patient care, a framework for score interpretation is required.

The National Institutes of Health (NIH) developed the Patient Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org) as part of a broader initiative to enhance the use and applicability of PROs for self-assessment of various domains of physical, mental, and social health. PROMIS measures were developed using a mixed methods approach, including item response theory (IRT), to develop reliable and valid tools. The PROMIS measures were developed for use across a range of disease conditions to assess both child and adult self-reported health. The measures are “generic” rather than disease specific, allowing for use and comparison across medical conditions [7, 8]. As of 2016, PROMIS pediatric measures are available to assess a wide range of domains of health including: Anxiety/Fear, Cognition, Depression/Sadness, Fatigue, Motor Function, Pain, Physical Function, Positive Affect/Well-being, Stress, Relationships/Social Support [9–15]. A complete and current list of instruments can be found at www.healthmeasures.net.

Despite their strong psychometric properties, the lack of an empirically established framework to interpret PROMIS scores in a clinically meaningful way impedes their use. Comparisons to normative general or clinical populations or another calibration sample provides a reference point but does not provide meaningful clinical information for decision-making. It also does not associate severity with scores. In addition, clinical meaningfulness of specific magnitudes of score change is needed. Furthermore, use of PROMIS in evaluation of treatment effectiveness requires establishment of minimally important differences (MID) in change scores [16, 17].

Standard setting for symptom severity is a new technique recently pioneered in oncology and multiple sclerosis (MS) populations by Cella et al. [18] in which cut-scores are developed for multi-item IRT-based PRO measures in health outcomes assessment. The method is based on modifications of the “bookmarking” method, a well-established approach for standard setting in fields of educational and psychological testing using IRT-calibrated items [19]. Standard setting methods are used to determine “valid and defensible” cut-scores for decisions associated with “high-stakes” consequences. The parallels between bookmarking in educational and in health research have been described [18]. Cella and colleagues modified this technique by ordering multi-item vignettes representing health symptoms and establishing cut offs according to severity of problems a person might experience [20].

In this study, we applied the modified bookmarking standard setting method [18, 20] to identify clinically relevant cut points and classify scores for PROMIS pediatric measures of physical health (upper extremity function, pain, fatigue, and mobility) by severity (none, mild, moderate, severe) for children with JIA. Additionally, we conducted a novel exercise to ascertain the magnitude of change, or differences in scores, that were deemed clinically significant by stakeholders.

Methods

Procedure

We used the modified bookmarking method to identify clinical severity thresholds for four PROMIS pediatric measures (mobility, upper extremity function, pain interference, and fatigue). These were obtained separately from adolescent persons with JIA (pwJIA), parents of JIA patients, and clinicians who treat children with JIA. Meetings were facilitated by a PhD level psychologist or psychometrician. Clinical vignettes were created to reflect likely item responses across a range of symptom severity (as described below). Panelists rank ordered vignettes by perceived symptom severity, and identified adjacent vignettes considered to represent upper and lower boundaries separating category cut points (i.e., no problems/mild problems, mild/moderate, moderate/severe). Cut-scores were defined as the mean score for boundary vignettes.

Developing Score-Level Vignettes

Measures—The PROMIS pediatric measures of mobility, upper extremity (UE) function, fatigue, and pain interference were selected because they quantify physical function and symptoms that can have a significant impact on daily functioning and well-being and are known to have independent effects on quality of life (QOL) of persons with JIA [21]. All PROMIS measures are calibrated using the graded response item response theory model. Details of the calibration, development and evaluation of these measures have been published elsewhere [7, 10, 12, 15].

All PROMIS pediatric measures use “The past 7 days…” as the reference period. For PROMIS mobility and UE function, the 5 response options ranged from “with no trouble” to “not able to do”, coded from 4 to 0, respectively. For fatigue and pain interference, the 5 response options ranged from “never” to almost always”, coded from 0 to 4, respectively. All PROMIS scores are reported on a T-score metric (mean = 50; SD = 10) and higher scores indicate more of the measured domain. For the UE function and mobility measures, a higher score reflects higher function. For the fatigue and pain interference measures, a higher score indicates worse fatigue or greater impact of pain.

Vignettes were based on most likely item responses at different levels on the T-score metric [mean = 50; SD = 10]. We selected T-score locations that spanned the range of scores measured by the short forms for each domain. In practice, different domains have different T-Score ranges, such that there are different highest and lowest possible T-Scores for each of the PROMIS measures. Clinical vignettes were created representing scores at 0.5 standard deviation intervals (i.e., 5 points of the T-Score metric) (see Figure 1). Because the T-Score ranges differed for each domain, the number of vignettes for each domain also differed. Specifically, we selected eleven T-Score locations for fatigue (corresponding to eleven clinical vignettes), ranging from 27.5 to 82.5. For pain interference, there were eight T-Score locations, ranging from 42.5 to 77.5. For mobility, there were ten T-Score locations, ranging from 7.5 to 57.5. Finally, for UE function, eight T-score locations were identified, ranging from 7.5 to 42.5. Cut scores for different severity levels were assigned the value of the mean of the upper and lower cut-scores delimiting the vignettes. The location of vignettes was chosen so that the mean would be an integer value, for ease of use.

Identifying Predicted Item Responses

The most likely item response for every item at every T-score location (in half SD increments) was identified. This was accomplished using the IRT parameters provided by the PROMIS Assessment Center (https://www.assessmentcenter.net) and an R-program (R Core Team, http://www.R-project.org) written to identify most likely responses, based on item parameters, for every target location (see Appendix B). Figure 2 shows the results for a subset of PROMIS Pain Interference items.

Creating Clinical Vignettes

For each vignette, we selected five items for every target location (T-score) on the measurement continuum for a given measure. Vignettes were made more realistic by assigning a first name selected from popular names for children in the US. Due to JIA affecting predominantly females, the majority of the vignettes described female children with JIA. See Appendix A for all vignettes used in the study. We used a wide range of PROMIS items during vignette construction to prevent panelists from comparing vignettes' items side by side and determining severity by comparing responses to the same items across vignettes. The intention was that panelists would review the vignettes and develop a mental picture of each person depicted and make a determination of how much the individual would be impacted by a symptom or outcome relative to those represented by the other clinical vignettes. Each vignette was printed on a card so that they could be physically laid out in front of panelists and rank ordered by severity. Cards were color coded for each domain of health.

Participants for Expert Panels—We recruited participants for three expert panels. A patient expert panel was comprised of persons with JIA (pwJIA) between ages of 15 to 20; a parent expert panel was comprised of parents of children with JIA ages 5 and older (parents); and a clinician expert panel was comprised of clinicians with expertise in pediatric rheumatology and treatment of JIA (clinicians). Although PROMIS pediatric measures are suitable for self-report starting at the age of 8, recruitment of pwJIA for the bookmarking exercise was limited to adolescent patients aged 13 and older (up to age 21) due to the cognitive requirements (e.g., of completing exercises at home and ranking them by severity). Furthermore, younger children tend to participate less in group discussion when paired with older children [22]. In contrast to the patient panel, parents of children 5 and older were eligible for the panel given that PROMIS proxy-report measures serve children ages 5 and over.

PwJIA were recruited from the clinic population at a large Midwestern children's hospital rheumatology clinic; parents were recruited from the same clinic, as well as from parent participant members in a JIA quality improvement network (https://PR-COIN.org). Clinicians were recruited via emails sent to North American subject matter experts in the field of pediatric rheumatology who were experienced in clinical care of JIA. All procedures were approved by the Institutional Review Board, and panelists completed the informed consent process. Panelists were compensated $200 for participation in pre-work exercises and a day-long meeting.

Pre-Workshop Procedures

To increase familiarity with the PROMIS measures, patient panelists were mailed the PROMIS pediatric short forms for each domain and asked to complete them prior to the workshop. Parents received and completed the corresponding PROMIS proxy-reports. Clinical providers received and were asked to complete the corresponding PROMIS pediatric proxy-report short forms while thinking of one of their patients. Panelists were instructed to order the vignettes by severity of the symptom. Data on vignette order was submitted to the research project team for compilation and presentation at the meeting.

In-Person Workshop Procedures

In May 2014, we conducted one-day concurrent expert panel meetings with patients and parents in separate exercises. Another one-day meeting was held with clinical providers in July 2014. Panel meetings began with introductions and a warm-up exercise intended to familiarize panelists with the bookmarking method.

After a warm-up exercise, we proceeded to the clinical vignettes. The procedures for each domain were identical. Panelists were presented with cards that duplicated the clinical vignettes they had previously received by mail and had ordered by perceived severity in advance of the workshop. They were also given four paper bookmarks. We began by presenting on the projection screen the order that panelists had assigned to the vignettes during their pre-work and compared this ordering to the actual severity ordering of the vignettes (based on T-scores). Panelists were not identified in this presentation. One purpose of this was to show the panelists that their average rankings successfully approximated the mathematical rankings of the vignettes (based on the T-score locations). Our intention was to demonstrate to panelists the validity of their judgments.

Next, panelists were asked to place the vignettes in front of them, ordering them from least to most severe based on their T-Score locations. Next, working individually, they placed a “bookmark” between the first vignette they judged as having mild problems for that health domain and the vignette that was just below it in order. In successive steps, and continuing to work individually, they placed bookmarks at thresholds between mild and moderate problems and between moderate and severe problems for the domain (Figure 3).

Results were collected and quickly summarized by a recorder. After this step was completed, thresholds for levels of severity were calculated (i.e., mean locations of the two vignettes adjacent to the bookmark location). Discussion amongst the panelists was then conducted by the facilitator, in which each panelist was able to discuss the rationale behind the placement of the bookmark. Discussion continued until consensus was obtained amongst panelists on the placement of the bookmarks and location of severity thresholds. This information was then used by one member of the team to produce “consequential validity” results as described below.

As a brief validity check, we then used the score distribution from the PROMIS item bank longitudinal study collected with 121 JIA patients [21], to show the panelists the percentages of the clinical sample that would be classified as having “no problems”, “mild problems”, “moderate problems”, and “severe problems” based on the location of the panelists’ thresholds. We then asked panelists whether they would have expected a different severity distribution and, if so, we allowed panelists to adjust their bookmarks.

Establishing MIDs

To define MIDs, panelists were given more extensive score reports for selected vignettes. The reports showed most likely answers to all items in the PROMIS domain item bank, not just the five items included in the vignette (see Figure 4). Panelists were instructed to review the items in the bank and judge which items on which a score change would be a particularly important indicator of improvement. Next, they were asked to respond to each of the items in a way that would represent “just enough improvement to make a difference” relative to the response given for the hypothetical patient. They were told that they were not “required” to show improvement in every item, nor were they required to make changes of only one response category. Panelists worked independently in scoring the items to represent change. After the workshop, the responses on the “improved” score report were used to calculate the associated T-scores using the freeware program, Firestar [23]. Thus, for each participant and each hypothetical score report, we had the anchoring T-score and a new T-score based on each panelists’ views of what constituted enough improvement to make a difference. The difference between these scores represented the amount of improvement the panelist judged to be “just enough improvement to make a difference”. The mean of these differences scores across panelists and vignettes were used as an estimate of the minimally important difference for a given measure.

Results

Participant Characteristics

Panels consisted of four pwJIA, five parents of pw JIA, and seven clinicians who treat JIA. The pwJIA panelists were all female and Caucasian and reported no other chronic conditions; they ranged in age from 15–20 years. Parent panelists were all Caucasian females, with at least some college education, and mothers of pwJIA aged 13 to 20. There were three parent/child dyads amongst the panelists. Clinicians were three male and four female pediatric rheumatology specialists from three academic medical practices in the Midwest (six medical doctors and one nurse practitioner) with experience ranging from 5 to 35 years post training completion. Details on the panelists are presented in Table 1. Table 2 presents the demographic and disease characteristics of the clinical sample. The majority of patients were female (71.1%) with polyarticular arthritis (63.6%). Data from the clinical sample was used to understand consequential validity of the cut-point selections.

Cut Scores by Domain and Panel

Table 3 summarizes the consensus-derived cut-scores. When comparing results across measures, there was more agreement on the cut-points for fatigue and UE Function across panels than for the mobility and pain interference. For fatigue, all panels selected cut-points within 0.5 SDs of each other. Clinicians tended to select lower cut-points than did parents and pwJIA. For example, clinicians judged fatigue to be a severe problem at a lower score than did parents or pwJIA. For UE Function, the cut-points were identical among panels for moderate and severe categories, and within 0.5 SD for the mild category; in this case pwJIA selected a higher cut-point than parents and clinicians, pwJIA considered a lower level of function to be a non-impaired (normal) state.

The cut-scores set by the expert panels for the mobility and pain interference domains showed larger differences. For pain interference's mild and moderate categories, pwJIA set the highest cut points relative to both parents and clinicians. To pwJIA, a relatively higher degree of pain interference was characterized as normal or mild. Clinicians saw pain interference as more problematic at lower levels. PwJIAs' and clinicians' cut points differed by the largest amount (~1 SD). Parents' ratings fell between. For mobility, pwJIA and clinicians set similar cut points, within 0.5 SD. The greater divergence in cut points was between parents and pwJIA. These cut-points differed by approximately 1 SD. For mobility, as with UE Function, pwJIA considered a lower level of function to be a non-impaired (normal) state (40), while parents set a higher cut-score (>52.5).

We considered the proportion of patients in the clinical validation sample that would fall into each severity category according to the panelists' cut-scores (see Figures 5a–5d). Based on these, about 50% would be classified as having mild to moderate problems with fatigue. A small proportion (3–4%) would be classified as having severe problems. With respect to UE Function, the vast majority (87–95%) would be considered to have no impairment. Consistent with the disparity in cut-scores for pain interference and mobility, there was variability in classification of patients depending on the panel group. Clinicians reported pain interference to be more problematic in the sample (60% of patients mild or greater pain interference), than did either parents (46%) or pwJIA (31%), respectively. Considering mobility, parents' cut points classified 61% of patients as having mild or moderate problems with mobility compared to clinician rating 47% and pwJIA rating 32% accordingly. None of the patients would have been considered to have severe mobility problems.

After reviewing the clinic population's distribution of severity according to the cut-scores established by the panelists during the exercise (i.e., consequential validity of the cut scores), no panelists altered their selected cut-scores.

Establishing MIDs by Domain and Panel

Estimates of MID varied according to severity classification of the symptom and the domain (see Table 3). Parents tended to estimate larger MIDs compared to other panelists (range: 1.3–12.7). Clinicians tended to estimate the lowest MIDs of all (range: 0.01–5.3). All panelists tended to estimate higher MIDs for fatigue (range: 3.0–9.4 for severe; 3.0–4.8 for mild) and pain interference (range: 5.3–12.7 for severe; 2.1–5.5 for mild), and these varied more by severity classification (larger MID for severe and smaller for mild category, than they did for mobility (range: 2.2–4.4 for severe, 1.6–5.4 for mild) and UE function (range: 1.8–2.8 for severe; 1.5–3.5 for mild). In addition, the differences among MIDs selected by the different panels was much greater for fatigue and pain interference and less variable in mobility and UE function.

Discussion

We used a modified educational standard setting method to estimate clinically relevant cut-scores to classify severity based on PROMIS measures relevant to pwJIA: mobility, upper extremity function, fatigue, and pain interference. Also, we used a novel extension of this methodology to estimate minimally important improvement values from the perspective of patients, parents, and clinicians.

Cut Scores

We found substantial congruence in estimated cut-scores across panel groups for UE function and fatigue, and some divergence for mobility and pain interference. When there was divergence, patients tended to place their cut scores at highest dysfunction, and parents the lowest dysfunction for severity classifications. This is consistent with a previous bookmarking study completed with persons who have MS (pwMS) where pwMS set higher cut-scores for higher levels of symptom severity than clinicians [18]. It is also consistent with other reports of incongruence in symptom ratings of patients and clinicians [24, 25].

The close agreement among panelists for fatigue cut-points may reflect that fatigue is a common human experience, disease related or not, and therefore evaluated more similarly across the individuals. Relatedly, divergence in cut-scores set for pain interference and mobility may be due to these being more "lived" experiences that parents and clinicians may be less familiar with personally. Perhaps pwJIA have adapted to the condition, and have higher tolerance for problems in these areas than when parents and clinicians can imagine living with the problems. They have a different internal standard as a consequence of living with JIA [26]. The expression from one of the pwJIA panelists that "it's [pain interference] just normal to me", may be indicative of such adaptation. In contrast, group discussion during the workshop revealed that the parents were comparing the affected child to peers, other children, or their "normal" baseline before becoming ill. Interestingly, clinicians assigned higher severity levels to lower pain interference scores, compared to both parents and pwJIA.

These differences in perspective beg the question of how to reconcile perceptions in making treatment decisions based on PRO scores. This type of incongruity has also been observed in self- vs. parent-proxy report score of various domains of health when using PROMIS and other measures, though with typically more congruence among physical health domains (such as were studied here) compared to emotional distress and peer relationship domains [27, 28]. Cook et al [18] recommended using cut-scores derived from persons with the condition of study, which were higher than set by clinicians. They argued for these as conservative thresholds for PRO alert notification to clinicians. For JIA, however, the choice is more complicated. There are substantial potential consequences of treatment recommendations based on cut score locations. Such recommendations may need to be influenced by the understanding that children and adolescents have not finished developing cognitively, which may limit medical decision making readiness [29]. Regardless, these differences highlight that parents and pediatric patients may have varying perspectives about severity and both should be considered in treatment decisions.

Consequential Validity

Our study considered the validity of the severity distribution by characterizing a clinical sample into severity levels based on the derived cut-scores and having panelists evaluate the face validity of this distribution from their perspective (patient, parent, or stakeholder). Based on the application of the cut scores, at least 50% of the clinical sample would be classified as having "problems" (mild to severe problems) with fatigue, mobility, and pain interference but only 13% with UE function. Given the clinical sample was largely a mix of patients already receiving treatment this may be a valid estimation. The panelists did not make any changes based on this presentation of data which was intended to explore the consequential validity of the cut-scores location. In hindsight it may have been unrealistic for parents and pwJIA to judge this consequence since they would be unlikely to be familiar with the distribution of symptoms and function in a general clinic sample. In future applications, other anchors for consequential validity should be explored. For example, it may be more relevant to present scenarios linked to discrete medical decisions based on severity category.

MIDs

The size of change scores deemed clinically relevant varied by severity, domain, and panel. The relatively higher magnitude of the MID at the more severe end of pain interference and fatigue compared to the mild end, may reflect less tolerance for high levels of pain and fatigue than impaired function amongst patients [30]. Clinicians tended to define smaller changes as important relative to others. This may reflect clinicians' familiarity with MID in the context of clinical trials, which often evaluate relatively small changes [31]. Parents tended to require the largest amount of change to reflect important differences. Differences among the panels may result from the instruction's ambiguity, especially vis-à-vis context. We asked panelists to indicate how much change would be "just enough to make a difference," but we did not specify the context, (e.g., requiring use of a new medication). Future investigators designing bookmarking studies for establishing thresholds for severity classifications or estimating clinically important change should consider ways of grounding panelists' judgments in clear and relevant contexts.

Advantages to Current Approach

There is growing awareness of the value of co-production in healthcare [32] and the importance of incorporating the perspective of the person experiencing the treatment impact on what truly constitutes efficacy, over statistical determinations, particularly when PRO scores serve as triggers for clinical action [18]. Our qualitative approach to understanding cut-scores from the perspective of patient, parent and clinician stakeholders revealed important differences that warrant replication. This finding highlights the shortcoming of even the most sophisticated of statistically based approaches that may fail to appreciate the subjective clinical impact of a statistically derived cut score on a person's function [33]. Our novel approach to MID determination as an initial approach likewise requires refinement. However, the value of the qualitative interview on MID determination in understanding the amount of change and in which elements of function an individual patient identifies as important is relevant in real world application to shared decision making and assessment of treatment values and preferences. One could envision a clinical application that matches potential treatment benefit of a novel treatment with a personalized valuation of the likely required benefit. While the MID scale–judgment technique [34] took a similar approach of respondents reviewing items and identifying importance of change in item responses, the items and amount of change were pre-selected rather than actively selected by the respondent as in our study. Furthermore, the scale-judgment method seemed subject to respondent error as evidenced by wrong-direction judgments. While Thissen et al. elicited differences in MID from various stakeholder perspectives in fact their valuations were relatively similar [34] when contrasted with our study which showed more marked variability, although this may in part be a reflection of our relatively small sample size.

Limitations

Despite our study's strengths, it also has limitations. The panels were small and homogenous. Future research should use a larger more diverse sample (e.g., clinical characteristics, age, race/ethnicity, etc.) to evaluate the replicability and generalizability of our findings. Similarly, a sample with a large number of patient-parent dyads would allow a detailed examination of concordance. Also, in retrospect, we realized the instructions for MIDs would have been better with a clearer context. Another limitation is that the parent panel had a different facilitator than the clinician and patient panels, confounding differences across panels with facilitator. Finally, we only examined MIDs within the context of improvement of function in these domains. It is possible that different MID scores would be found if panelists were asked to consider worsening of function. Indeed, previous work has demonstrated that MIDs do tend to differ when presented as worsening of function versus improving [35].

Conclusions

This study provides evidence to help with interpretation of PROMIS scores in a clinical setting with patients with JIA. Our investigation uncovered differences in perspectives on symptom severity and degree of change in a symptom that is meaningful between pwJIA, parents and clinicians. This is an area that warrants future study as it impacts approach to shared decision making around treatment interventions.

Acknowledgments

Funding: The Patient-Reported Outcomes Measurement Information System (PROMIS) is an NIH Roadmap initiative to develop a computerized system measuring PROs in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS II was funded by cooperative agreements with a Statistical Center (Northwestern University, PI: David Cella, PhD, 1U54AR057951), a Technology Center (Northwestern University, PI: Richard C. Gershon, PhD, 1U54AR057943), a Network Center (American Institutes for Research, PI: Susan (San) D. Keller, PhD, 1U54AR057926) and thirteen Primary Research Sites which may include more than one institution (State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, 1U01AR057948; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, 1U01AR057954; University of Washington, Seattle, PIs: Dagmar Amtmann, PhD and Karon Cook, PhD, 1U01AR052171; University of North Carolina, Chapel Hill, PI: Darren A. DeWalt, MD, MPH, 2U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, 1U01AR057956; Stanford University, PI: James F. Fries, MD, 2U01AR052158; Boston University, PIs: Stephen M. Haley, PhD and David Scott Tulsky, PhD (University of Michigan, Ann Arbor), 1U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD and Brennan Spiegel, MD, MSHS, 1U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, 2U01AR052155; Georgetown University, PIs: Carol. M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle) and Arnold L. Potosky, PhD, U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan DeWitt, MD, MSCE, 17 1U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, 1U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, 2U01AR052186). NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD, Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein, MD, Ellen Werner, PhD and James Witter, MD, PhD.

We are grateful for the contributions of the study participants, patients, parents, and clinicians who shared their time and insights to advance this work. We acknowledge Dr. Ryoungsun Park http://coe.wayne.edu/profile/fy3504 who developed the R-code for this work as part of a consultancy agreement with Northwestern University.

References

1. Barth S, et al. Long-Term Health-Related Quality of Life in German Patients with Juvenile Idiopathic Arthritis in Comparison to German General Population. PLoS One. 2016; 11(4):e0153267. [PubMed: 27115139]
2. Armbrust W, et al. Fatigue in patients with juvenile idiopathic arthritis: A systematic review of the literature. Semin Arthritis Rheum. 2016; 45(5):587–95. [PubMed: 26656031]
3. Hoeksma AF, et al. High prevalence of hand- and wrist-related symptoms, impairments, activity limitations and participation restrictions in children with juvenile idiopathic arthritis. J Rehabil Med. 2014; 46(10):991–6. [PubMed: 25188280]
4. Moorthy LN, et al. Burden of childhood-onset arthritis. Pediatr Rheumatol Online J. 2010; 8:20. [PubMed: 20615240]
5. Broderick J, et al. Advances in patient reported outcomes: The NIH PROMIS measures. eGEMs. 2013; 1:2327–9214.1015. Article 12. 10.
6. Jensen RE, et al. The role of technical advances in the adoption and integration of patient-reported outcomes in clinical care. Med Care. 2015; 53(2):153–9. [PubMed: 25588135]
7. Reeve BB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007; 45(5 Suppl 1):S22–31. [PubMed: 17443115]

8. Witter JP. The Promise of Patient-Reported Outcomes Measurement Information System-Turning Theory into Reality: A Uniform Approach to Patient-Reported Outcomes Across Rheumatic Diseases. Rheum Dis Clin North Am. 2016; 42(2):377–94. [PubMed: 27133496]
9. Irwin DE, et al. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. Qual Life Res. 2010; 19(4):595–607. [PubMed: 20213516]
10. Varni JW, et al. PROMIS Pediatric Pain Interference Scale: an item response theory analysis of the pediatric pain item bank. J Pain. 2010; 11(11):1109–19. [PubMed: 20627819]
11. Irwin DE, et al. PROMIS Pediatric Anger Scale: an item response theory analysis. Qual Life Res. 2012; 21(4):697–706. [PubMed: 21785833]
12. DeWitt EM, et al. Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: built using item response theory. J Clin Epidemiol. 2011; 64(7):794–804. [PubMed: 21292444]
13. Dewalt DA, et al. PROMIS Pediatric Peer Relationships Scale: development of a peer relationships item bank as part of social health measurement. Health Psychol. 2013; 32(10):1093–103. [PubMed: 23772887]
14. Varni JW, et al. PROMIS(R) Parent Proxy Report Scales for children ages 5–7 years: an item response theory analysis of differential item functioning across age groups. Qual Life Res. 2014; 23(1):349–61. [PubMed: 23740167]
15. Lai JS, et al. Development and psychometric properties of the PROMIS((R)) pediatric fatigue item banks. Qual Life Res. 2013; 22(9):2417–27. [PubMed: 23378106]
16. Guyatt GH, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002; 77(4):371–83. [PubMed: 11936935]
17. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987; 40(2):171–8. [PubMed: 3818871]
18. Cook KF, et al. Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. Qual Life Res. 2015; 24(3):575–89. [PubMed: 25148759]
19. Zieky, MJ., Perie, M., Livingston, SA. Cutscores: A manual for setting standards of performance on educational and occupational tests. Educational Testing Service; 2008.
20. Cella D, et al. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. Qual Life Res. 2014; 23(10):2651–61. [PubMed: 24938431]
21. Huang B, et al. ACR Criteria, Providers' Global Rating of Change and Role of Patient Self-Report in Evaluating Change in Disease Over Time: A Patient Reported Outcomes Measurement Information System Study. Arthritis Rheum. 2012; 64 Supplement(10)
22. Jacobson CJ Jr, et al. Qualitative Evaluation of Pediatric Pain Behavior, Quality, and Intensity Item Candidates and the PROMIS Pain Domain Framework in Children With Chronic Pain. J Pain. 2015; 16(12):1243–55. [PubMed: 26335990]
23. Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. Applied Psychological Measurement. 2009; 33(8)
24. Radowsky JS, et al. Pain ratings by patients and their providers of radionucleotide injection for breast cancer lymphatic mapping. Pain Med. 2012; 13(5):670–6. [PubMed: 22536858]
25. Basch E, et al. Patient versus clinician symptom reporting using the National Cancer Institute Common Terminology Criteria for Adverse Events: results of a questionnaire-based study. Lancet Oncol. 2006; 7(11):903–9. [PubMed: 17081915]
26. Brossart DF, Clay DL, Willson VL. Methodological and statistical considerations for threats to internal validity in pediatric outcome data: response shift in self-report outcomes. J Pediatr Psychol. 2002; 27(1):97–107. [PubMed: 11726684]
27. Varni JW, et al. Item-level informant discrepancies between children and their parents on the PROMIS((R)) pediatric scales. Qual Life Res. 2015; 24(8):1921–37. [PubMed: 25560776]
28. Lal SD, et al. Agreement between proxy and adolescent assessment of disability, pain, and well-being in juvenile idiopathic arthritis. J Pediatr. 2011; 158(2):307–12. [PubMed: 20869068]
29. Lipstein EA, et al. "I'm the one taking it": adolescent participation in chronic disease treatment decisions. JAdolesc Health. 2013; 53(2):253–9. [PubMed: 23561895]

30. Guzman J, et al. What matters most for patients, parents, and clinicians in the course of juvenile idiopathic arthritis? A qualitative study. J Rheumatol. 2014; 41(11):2260–9. [PubMed: 25225279]

31. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences. Analyses in 8931 patients with rheumatoid arthritis. J Rheumatol. 2005; 32(4):583–9. [PubMed: 15801011]

32. Batalden M, et al. Coproduction of healthcare service. BMJ Qual Saf. 2016; 25(7):509–17.

33. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991; 59(1):12–9. [PubMed: 2002127]

34. Thissen D, et al. Estimating minimally important difference (MID) in PROMIS pediatric measures using the scale-judgment method. Qual Life Res. 2016; 25(1):13–23. [PubMed: 26118768]

35. Brunner HI, et al. Minimal clinically important differences of the childhood health assessment questionnaire. J Rheumatol. 2005; 32(1):150–61. [PubMed: 15630741]

Appendix A: Clinical Vignettes

(Note: *T* score locations are provided, but were not printed on participant copies)

Fatigue

Anne’s Fatigue (T Score = 27.5)

In the last 7 days, Anne never got tired easily and never had trouble starting things because she was too tired. She felt that she was never so tired that it was hard for her to focus on work. She never found it hard f to get out of bed in the morning because she was too tired and she never felt too tired to do sports or exercise.

In summary, Anne reports:

- Never getting tired easily
- She never had trouble starting things because she was too tired
- Never feeling so tired that it was hard for her to focus on work
- It was never hard to get out of bed in the morning because of being too tired
- Never feeling too tired to do sports or exercise

Taylor’s Fatigue (T Score = 37.5)

In the last 7 days, Taylor sometimes felt tired. She was never too tired to go up and down a lot of stairs and she never had trouble finishing things because she was too tired. It was never hard for her to play or go out with her friends as much as she likes, and being tired never made it hard for her to keep up with her schoolwork.

In summary, Taylor reports:

- Sometimes feeling tired
- Never being too tired to go up and down a lot of stairs
- She never had trouble finishing things because she was too tired
- Never being so tired that it was hard for her to play or go out with her friends as much as she would like

- It was never hard for her to keep up with her schoolwork due to being tired

David’s Fatigue (T Score = 42.5)

In the last 7 days, David sometimes felt tired but almost never got tired easily. He was never too tired to do things outside or to enjoy the things that he likes to do. Similarly, he was never so tired that it was hard for him to focus on his work.

In summary, David reports:

- Sometimes feeling tired
- He almost never got tired easily
- He was never too tired to do things outside
- He was never too tired to enjoy things that he likes to do
- It was never hard to focus on work due to being tired

Zoe’s Fatigue (T Score = 47.5)

In the last 7 days, Zoe was never too tired to read or to watch television. She almost never felt weak, but sometimes felt more tired than usual when she woke up in the mornings. She also mentioned that it was sometimes hard for her to get out of bed in the morning because she was too tired.

In summary, Zoe reports:

- She was never too tired to read
- Never being too tired to watch television
- Almost never feeling weak
- Sometimes she felt more tired than usual when waking up in the morning
- Sometimes it was hard to get out of bed in the morning because she was too tired

Katie’s Fatigue (T Score = 52.5)

In the last 7 days, Katie sometimes got tired easily. Being tired never made it hard for her to keep up with her schoolwork, and she almost never had trouble starting things because she was too tired. In addition, she was never too tired to take a bath or shower but sometimes she was so tired that it was hard for her to pay attention.

In summary, Katie reports:

- Sometimes she got tired easily
- Being tired never made it hard for her to keep up with her schoolwork
- She almost never had trouble starting things because of tiredness
- She was never too tired to take a bath or shower
- Sometimes she was so tired that it was hard for her to pay attention

Jessica’s Fatigue (T Score = 57.5)

In the last 7 days, Jessica sometimes was too tired to do things outside. Sometimes she had trouble finishing things because she was too tired, and being tired sometimes kept her from having fun. She was also sometimes too tired to read. She never felt too tired to eat.

In summary, Jessica reports:

- Sometimes she was too tired to do things outside
- Sometimes she had trouble finishing things because she was too tired
- Being tired sometimes kept her from having fun
- Sometimes she felt too tired to read
- She never felt too tired to eat

Nevaeh’s Fatigue (T Score = 62.5)

In the last 7 days, Nevaeh sometimes was too tired to go up and down a lot of stairs and sometimes she was too tired to do sports or exercise. While she was almost never too tired to go out with her family, sometimes she needed to sleep during the day. Being tired sometimes made it hard for her to keep up with her school work.

In summary, Nevaeh reports she:

- Sometimes she was too tired to go up and down a lot of stairs
- Sometimes she was too tired to do sports or exercise
- Almost never felt too tired to go out with her family
- Sometimes needed to sleep during the day
- Sometimes being tired made it hard to keep up with schoolwork

Grace’s Fatigue (T Score = 67.5)

In the last 7 days, Grace was sometimes too tired to do things outside. It was almost always hard for her to get out of bed in the morning and she often had trouble finishing things because she was too tired. She sometimes felt weak. Being tired sometimes made it hard for her to play or go out with her friends as much as she would like.

In summary, Grace reports she:

- Sometimes was too tired to do things outside
- Almost always found it hard to get out of bed in the morning because of tiredness
- Often had trouble finishing things because of tiredness
- Sometimes felt weak
- Sometimes because of being tired she found it hard to play or go out with friends as much as she would have liked

Isabel's Fatigue (T Score = 72.5)

In the last 7 days, Isabel almost always felt more tired than usual when she woke up in the morning. She often felt tired and sometimes felt too tired to spend time with her friends. Isabel was sometimes so tired that it was hard to focus on her work and she was almost always too tired to read.

In summary, Isabel reports she:

- Almost always felt more tired than usual when waking up in the morning
- Often felt tired
- Sometimes felt too tired to spend time with her friends
- Sometimes was so tired that is was hard to focus on work
- Almost always was too tired to read

Owen's Fatigue (T Score = 77.5)

In the last 7 days, Owen was sometimes too tired to eat and almost always too tired to do sports and exercise. He often had trouble starting things because he was too tired and felt that being tired almost always kept him from having fun. Owen almost always found being tired made it hard to play or go out with friends as much as he'd like.

In summary, Owen reports:

- Sometimes being too tired to eat
- Almost always being too tired to do sports and exercise
- He often had trouble starting things because of being tired
- Being tired almost always kept him from having fun
- It was almost always hard to play or go out with friends as much as he would have liked due to being tired

Leah's Fatigue (T Score = 82.5)

In the last 7 days, Leah was sometimes too tired to watch television, and almost always too tired to take a bath or shower. She was also almost always too tired to go out with her family, or enjoy the things that she liked to do. Being tired almost always made it hard for Leah to keep up with her schoolwork.

In summary, Leah reports she:

- Sometimes was too tired to watch television
- Almost always was too tired to take a bath or shower
- Almost always was too tired to go out with her family
- Almost always was too tired to enjoy the things that she likes to do
- Almost always had a hard time keeping up with her schoolwork

Upper Extremity Function

Ella’s Upper Extremity (T Score = 7.5)

In the last 7 days, Ella was not able to cut paper with scissors, nor was she able to lift a cup to drink. She was unable to open her clothing drawers and she could not take a bath without help. She also was not able to put on her shoes unassisted.

In summary, Ella reported she:

- Could not cut paper with scissors
- She was not able to lift a cup to drink
- Could not open her clothing drawers
- Almost always needed help with a bath
- Was not able to put on her shoes by herself

Mary Kate’s Upper Extremity (T Score =12.5)

In the last 7 days, Mary Kate had a lot of trouble writing with a pen or a pencil. She not able to zip up her clothes, nor was she able to unlock a door with a key. She was also not able to open a jar or pour a drink from a full pitcher by herself.

In summary, Mary Kate reported she:

- Could write with a pen or pencil only with a lot of trouble
- Was unable to zip up her clothes
- Could not unlock a door with a key
- Could not open a jar by herself
- Was not able to pour a drink from a full pitcher by herself

Chloe’s Upper Extremity (T Score = 17.5)

In the last 7 days, Chloe could put her shoes on by herself with a lot of trouble, but was not able to put on her socks. She could lift a cup to drink with some trouble. She was not able to open the rings in school binders nor was she able to open heavy doors.

In summary, Chloe reported she:

- Could put on her shoes by herself with a lot of trouble
- Was unable to put on her socks by myself
- Could lift a cup to drink, but with some trouble
- Was unable to open the rings in school binders
- Could not pull open heavy doors

Brianna’s Upper Extremity (T Score = 22.5)

In the last 7 days, Brianna had some trouble with opening jars. She had a little trouble with cutting paper with scissors, a little trouble with opening her clothing drawers, and a little trouble zipping up her clothes. However, she could not put toothpaste on her toothbrush by herself.

In summary, Brianna reported she:

- Could open a jar with some trouble
- Was able to cut paper with scissors with a little trouble
- Could open her clothing drawers with a little trouble
- Could zip up her clothes with a little trouble
- Was unable to put toothpaste on her toothbrush by herself

Emma’s Upper Extremity (T Score = 27.5)

In the last 7 days, Emma never needed help with a bath. She could dry her back with a towel with a little trouble, and with a little trouble she could put on her socks by herself. She had a little trouble unlocking a door with a key. She had no trouble with lifting a cup to drink.

In summary, Emma reports she:

- Never needed help with a bath
- Could dry her back with a towel with a little trouble
- Could put on socks with a little trouble
- Had a little trouble using a key to unlock a door
- Could lift a cup to drink with no trouble

Sophia’s Upper Extremity (T Score = 32.5)

In the last 7 days, Sophia has had a little trouble with opening the rings in her binders for school. She also had a little trouble with pouring a drink from a full pitcher. She had no trouble washing her face with a cloth, opening her clothing drawers, or writing with a pen or pencil.

In summary, Sophia reports she:

- Could open the rings in school binders with a little trouble
- Could pour a drink from a full pitcher with a little trouble
- Had no trouble washing her face with a cloth
- Had no trouble opening her clothing drawers
- Could write with a pen or pencil with no trouble

Allison's Upper Extremity (T Score = 37.5)

In the last 7 days, Allison could put on her shoes by herself, pull on and fasten her seat belt, and dial a phone with no trouble. She also could put toothpaste on her toothbrush by herself with no trouble. She had a little trouble with opening a jar.

In summary, Allison reports she:

- Could put on her shoes by herself with no trouble
- Had no trouble pulling on and fastening her seat belt
- Had no trouble dialing a phone
- Could put toothpaste on her toothbrush by herself without trouble
- Could open a jar with a little trouble

Hailey's Upper Extremity (T Score = 42.5)

In the last 7 days, Hailey could use a key to unlock a door with no trouble, and had no trouble pulling open heavy doors. She could brush her teeth by herself. She also could zip up her clothes and put on her socks by herself with no trouble.

In summary, Hailey reported she:

- Could unlock a door with a key with no trouble
- Could pull open heavy doors with no trouble
- Had no trouble with brushing her teeth by herself
- Could zip up her clothes with no trouble
- Could put on socks with no trouble

Pain Interference

Andrea's Pain (T Score = 42.5)

In the last 7 days, Andrea never found it hard to get along with other people or pay attention when she had pain. She never hurt a lot and never found it hard to walk one block when she had pain. It was also never hard for Andrea to stay standing when she had pain.

In summary, Andrea reports:

- It was never hard to get along with other people when she had pain
- It was never hard to pay attention when she had pain
- She never hurt a lot
- Never finding it hard to walk one block when she has pain
- It was never hard to stay standing when she has pain

Claire’s Pain (T Score = 47.5)

In the last 7 days, Claire sometimes found it hard to run when she was in pain. But her pain never caused her to feel angry. She never found it hard to have fun when in pain. She also never had trouble sleeping and never missed school when she had pain.

In summary, Claire reports:

- Sometimes finding it hard to run when she had pain
- Never feeling angry when she had pain
- It was never hard for her to have fun when she has pain
- She never had trouble sleeping when she has pain
- She never missed school when she has pain

Maya’s Pain (T Score = 52.5)

In the last 7 days, Maya never hurt all over her body, but sometimes had trouble sleeping when she had pain. She never found it hard to walk one block when in pain. Maya almost never had trouble doing schoolwork when she had pain, but sometimes found it hard to pay attention when she had pain.

In summary, Maya reports she:

- Never hurt all over her body
- Sometimes has trouble sleeping when in pain
- Never found it hard to walk one block when in pain
- Never has trouble doing schoolwork when in pain
- Sometimes has a hard time paying attention when in pain

Addison’s Pain (T Score = 57.5)

In the last 7 days, Addison found it sometimes hard to have fun when she had pain. It was never hard for her to remember things when she had pain. She sometimes felt angry and sometimes had trouble doing schoolwork when she had pain. She sometimes found it hard to stay standing when she had pain.

In summary, Addison reports:

- Sometimes it was hard to have fun when in pain
- It was never hard to remember things when in pain
- Sometimes feeling angry when she had pain
- Sometimes she had trouble doing schoolwork when in pain
- Sometimes it was hard to stay standing when she had pain

Jacob’s Pain (T Score = 62.5)

In the last 7 days, Jacob often found it hard to run when he had pain. He sometimes hurt all over his body and found that when he had pain he sometimes missed school. It was also sometimes hard for him to walk one block or remember things when he had pain.

In summary, Jacob reports:

- Often found it hard to run when in pain
- Sometimes hurt all over his body
- Sometimes missed school when having pain
- Sometimes it was hard for him to walk one block when he had pain
- Sometimes it was hard to remember things when in pain

Anna’s Pain (T Score = 67.5)

In the last 7 days, Anna almost always had trouble sleeping and often had trouble doing schoolwork when she was in pain. She sometimes hurt a lot and it was almost always hard for her to stay standing when she had pain. She sometimes felt it was hard to get along with other people when she had pain.

In summary, Anna reports she:

- Almost always had trouble sleeping when in pain
- Often had trouble doing schoolwork when in pain
- Sometimes hurt a lot
- Almost always had a hard time staying standing when she had pain
- Sometimes finds it hard to get along with other people when in pain

Julia’s Pain (T Score = 72.5)

In the last 7 days, Julia almost always felt angry when she had pain and almost always found it hard have fun. She often found it hard to remember things when she had pain. She sometimes missed school when in pain. It was almost always hard for her to walk one block when in pain.

In summary, Julia reports she:

- Almost always feels angry when in pain
- Almost always finds it hard to have fun when in pain
- Often finds it hard to remember things when in pain.
- Sometimes misses school when in pain
- Almost always finds it hard to walk one block when in pain

Kristen's Pain (T Score = 77.5)

In the last 7 days, Kristen almost always hurt a lot and almost always hurt all over her body. She often found it hard to remember things when she had pain. Kristen also reported that it was almost always hard for to run and get along with other people she had pain.

In summary, Kristen reports she:

- Almost always hurts a lot
- Almost always hurts all over her body
- Often has a hard time remembering things when in pain
- Almost always finds it hard to run when in pain
- Almost always has a hard getting along with other people when she has pain

Mobility

Lily's Mobility (T Score =7.5)

In the past 7 days, Lily was unable to move her legs. She could not get into bed by herself, nor stand up by herself. She was unable to turn her head all the way to the side and could not bend over to pick something up.

In summary, Lily reports she:

- Could not move her legs
- Was unable to get into bed by herself
- Was unable to stand up by herself
- Could not turn her head all the way to the side
- Was unable to bend over to pick something up

Kylie's Mobility (T Score = 12.5)

In the past 7 days, Kylie was unable to keep up with other kids her age when they played together. With a lot of trouble she could go up one step. She was unable to stand on tiptoes, and could not walk more than one block. She could not get in and out of a car.

In summary, Kylie reports that she:

- Could not keep up when playing with other kids
- Could go up one step, but with a lot of trouble
- Could not stand on her tiptoes
- Could not walk more than one block
- Was unable to get in and out of a car

Madison's Mobility (T Score = 17.5)

In the past 7 days, Madison was able to get out of bed by herself with a little trouble. While she could bend over to pick something up with some trouble, she found she was not able to get up from the floor. Madison almost always used a wheelchair to get around, and she was physically unable to do the activities she enjoys the most.

In summary, Madison reports she:

- Could get out of bed by herself with a little trouble
- Could bend over to pick something up with some trouble
- Was unable to get up from the floor
- Almost always used a wheelchair to get around
- Was not physically able to do the activities she enjoys most

Chase's Mobility (T Score = 22.5)

In the past 7 days, Chase had a lot of trouble with moving his legs. However, he was able to go up one step with a little trouble, and he could get up from a regular toilet with no trouble. Chase was unable to get down on to his knees without support and he could not do sports and exercise that other kids his age could do.

In summary, Chase reports he:

- Could move his legs with a lot of trouble
- Could go up one step with a little trouble
- Was able to get up from a regular toilet with no trouble
- Could not get down on his knees without holding onto something
- Was unable to do sports and exercise other kids his age could do

Emily's Mobility (T Score = 27.5)

In the past 7 days, Emily found that she was unable to ride a bike. With some trouble she could walk more than one block and stand up on her tip toes. Emily could get in and out of a car with a little trouble and never needed to use a walker, cane or crutches to get around.

In summary, Emily reports she:

- Was unable to ride a bike
- With some trouble she could walk more than one block
- With some trouble she could stand up on her tip toes
- She could get in and out of a car with a little trouble
- Never needed to use a walker, cane or crutches to get around

Evelyn’s Mobility (T Score = 32.5)

In the past 7 days, Evelyn reports she could turn her head all the way to the side and get into bed by herself with no trouble. With a little trouble Evelyn could get down on her knees unassisted, or walk up stairs without holding onto anything. However, she had some trouble with running a mile.

In summary, Evelyn reports she:

- With no trouble could turn her head all the way to the side
- With no trouble could get into bed by herself
- With a little trouble could get down on her knees without holding onto something
- With a little trouble could walk up stairs without holding onto anything
- Could run a mile with some trouble

Olivia’s Mobility (T Score = 37.5)

In the past 7 days, Olivia had a little trouble with physically doing the activities she enjoys most. She found it was no trouble to stand up by herself and had no trouble moving her legs. She also had no trouble carrying her books in her backpack. Olivia had some trouble with doing sports and exercise that other kids her age could do.

In summary, Olivia reports she:

- With a little trouble she could physically do the activities she enjoys most
- Could stand up by herself with no trouble
- Could move her legs with no trouble
- Carried her books in her backpack with no trouble
- With some trouble could do sports and exercise that other kids her age could do

Lauren’s Mobility (T Score = 42.5)

In the past 7 days, Lauren found that she could get out of bed with no trouble. It was also no trouble for her to get up from the floor or to bend over to pick something up. She could ride a bike with no trouble; however she had a little trouble keeping up when she played with the other kids.

In summary, Lauren reports she:

- Could get out of bed by herself with no trouble
- Had no trouble getting up from the floor
- Could bend over to pick something up from the floor with no trouble
- Could ride a bike with no trouble
- Had a little trouble keeping up when she played with the other kids

Samantha’s Mobility (T Score = 47.5)

In the past 7 days, Samantha found herself able to get in and out of a car with no trouble, and also had no trouble with going up one step, or standing on her tiptoes. It was no trouble for her to do sports and exercise other kids her age could do. However, she did have a little trouble with running a mile.

In summary, Samantha reports she:

- Could get in and out of a car with no trouble
- Could go up one step with no trouble
- Had no trouble standing on her tiptoes
- Could do sports and exercise that other kids her age can do, with no trouble
- Was able to run a mile with a little trouble

Caroline’s Mobility (T Score = 57.5)

In the past 7 days, Caroline found that she could keep up with the other kids when playing. She had no trouble walking up the stairs without holding on to anything, or getting down on her knees without holding on to something. She could run a mile with no trouble, and has been physically able to do the activities she enjoys the most.

In summary, Caroline reports she:

- Could keep up with the other kids when playing with no trouble
- Was able to walk up stairs without holding on to anything with no trouble
- Could get down on her knees without holding on to something with no trouble
- Had no trouble running a mile
- Had no trouble being physically able to do the activities she enjoys the most

Appendix B. R Code to Generate Most Likely Item Responses for a Given T-score Created by Dr. Ryoungsun Park, http://coe.wayne.edu/profile/fy3504

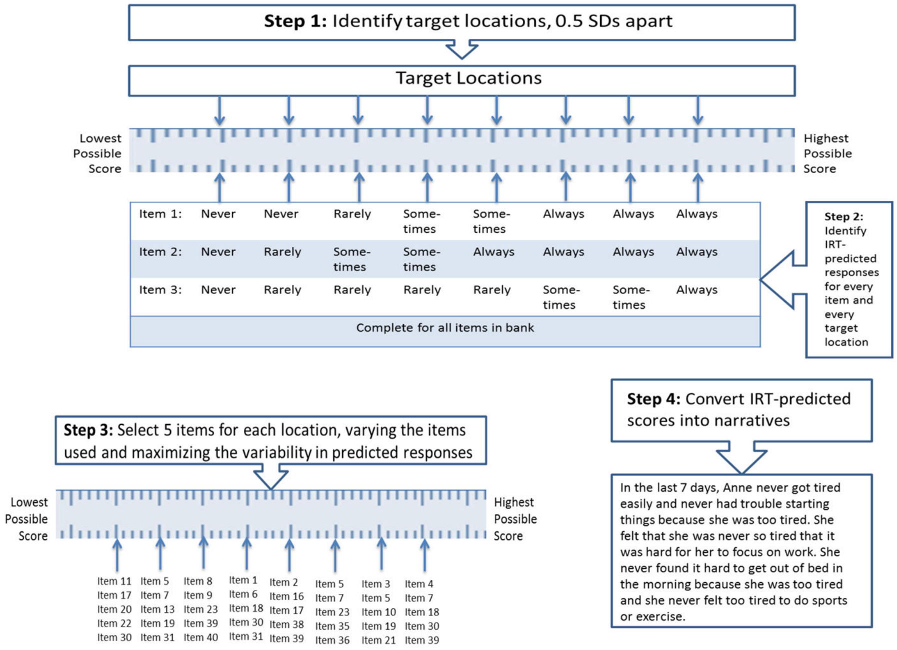
```
#############################################################################
###
# This code does the following:
#############################################################################
###
# 0. read item list file
# 1. read polytomous item parameters
# 2. calculate probabilities for each category for each items
# 3. obtain categories of maximum probability for an array of theta
# 4. save previous results to a file
# 5. plot item characteristic curves
```

```
##############################################################################
###
# item parameter file of following format
# this code strictly follows the file format, thus do not change the header
##############################################################################
###
#                 a                cb1           cb2          cb3
cb4 NCAT
#1  3.024324   0.17046739 0.7764859 1.790530 2.772405        5
#2  3.868153   0.52836169 1.0173187 1.676377 2.366108        5
#...
#30 2.296848 -0.83224545 0.2640290 1.438525 2.739884        5
##############################################################################
###
# item content file of following format
# this code strictly follows the file format, thus do not change the header
##############################################################################
###
#item content    order type ncat kept
#EDDEP02  I felt lonely even when I was with other people   1   O   5
               1
#EDDEP04  I felt worthless      2          O          5       1
##############################################################################
###
# variables that need programming
##############################################################################
###
domain<-”Depression”
ver<-1
#working directory that contains item list and parameter files
#output file will created under this working directory
workdir <- “C:/Users/[….]”
item.list.fn       <-paste(workdir , “Item_”,domain,“_v”,ver,“.csv”,sep=“”)
item.param.fn <-paste(workdir , “eq_par_”,domain,“_v”,ver,“.csv”,sep=“”)
outputfile        <-paste(workdir , “max_prob_cat_”,domain,“.csv”,sep=“”)
minTheta <- -2.0      #lower bound of theta
maxTheta <- 4.0       #larger bound of theta
increment <- 0.25       #spacing between two bounds
D <- 1
##############################################################################
###
# 0. read item list file
##############################################################################
###
```

```
res1 <- try( item.list<-read.table(item.list.fn,
sep=“,”,row.names=NULL,header=T) )
if(is.null(res1)){ stop(“file open error!, maybe wrong path or file does not
exist?”)
}
item.names<-as.character(item.list$Items)
item.content<-as.character(item.list$content)
item.ncat <-as.numeric(item.list$ncat)
maxCat <- max(item.ncat)
##############################################################################
###
# 1. read in item parameters
##############################################################################
###
res2 <- try( ipar<-read.table(item.param.fn,
sep=“,”,row.names=NULL,header=T) )
if(is.null(res2)){ stop(“file open error!, maybe wrong path or file does not
exist?”)
}
#simple error checking by comparing two files
maxcat_tmp <- max(as.numeric(ipar$NCAT))
if(maxCat != maxcat_tmp)    {
             stop(“max of categories in item parameter file and item file
are different”)
}
ni<-dim(ipar)[1]                                                        #num
of item
theta <- seq(minTheta ,maxTheta ,by=increment)   #theta
nq <- length(theta)                                                    #num
of quadratures
##############################################################################
###
# 2. calculate probabilities for each category for each items
##############################################################################
###
pp <- array(0, dim = c(ni, maxCat, nq))
calc.prob<-function() {
for(i in 1:ni) {
 a <- ipar[i,”a”]
 ncat <- ipar[i, “NCAT”]
 #make a vector for only categories (cb1,..) for item i
 cb <-unlist(ipar[i,paste(“cb”,1:(ncat-1),sep=““)])
 ps <- matrix(0,nrow =ncat+1, ncol =nq)
 ps[1,]<-1
 ps[ncat+1,]<-0
 for(k in 1:(ncat-1)) {
  ps[k+1,] <- 1/(1+exp(-D*a*(theta-cb[k])))
 }
 pp[i,1,] <- 1-ps[1,]
 pp[i,ncat,] <- ps[ncat,]
 for(k in 1:ncat) {
  pp[i,k,] <- ps[k,]-ps[k+1,]
 }
}
return(pp)
}
pp<-calc.prob()
##############################################################################
###
# 3. obtain categories of maximum probability for an array of theta
##############################################################################
###
max.prob.category.matrix<-matrix(0,ni,nq)
for (i in 1:ni) {
         for (t in 1:nq) {
                  max.prob.category.matrix[i,t]<-
which(pp[i,,t]==max(pp[i,,t]))
         }
}
##############################################################################
###
# 4. save result to a file
##############################################################################
###
out.matrix<-data.frame(item=item.names, content=item.content,
max.prob.category.matrix);
#change the names of variable from X?? to actual theta value
names(out.matrix)[3:(length(theta)+2)] <- theta
res3 <- try( write.csv(out.matrix,outputfile) )
if(!is.null(res3)){ stop(“file save error!, maybe wrong path or file already
open?”) }
##############################################################################
###
# 5. plot item characteristic curves
##############################################################################
###
plot.item.prob<-function () {
         for (i in 1:ni){
```

```
                  ncat<-ipar[i,”NCAT”]
               #this will create the plots in multiple windows and keep them
seperate in
R
               if(i%%12==1) {      #every 12 items
                 dev.new()                #create a new window
                 par(mfrow=c(3,4)); #set parameter for newly created window
               }

plot(theta,seq(0,1,length=length(theta)),type=“n”,xlab=“Theta”,ylab=“Probabil
it
y”,main=item.names[i]);
               for (k in 1:ncat){
                        lines(theta,pp[i,k,],lty=k);
               }
               legend(min(theta),
1,legend=as.character(1:ncat),lty=1:ncat,cex=0.5);
         }
}
plot.item.prob()
```


Step 1: Identify target locations, 0.5 SDs apart
Target Locations
Lowest Possible Score
Highest Possible Score
Step 2: Identify IRT-predicted responses for every item and every target location
Step 3: Select 5 items for each location, varying the items used and maximizing the variability in predicted responses
Lowest Possible Score
Highest Possible Score
Step 4: Convert IRT-predicted scores into narratives

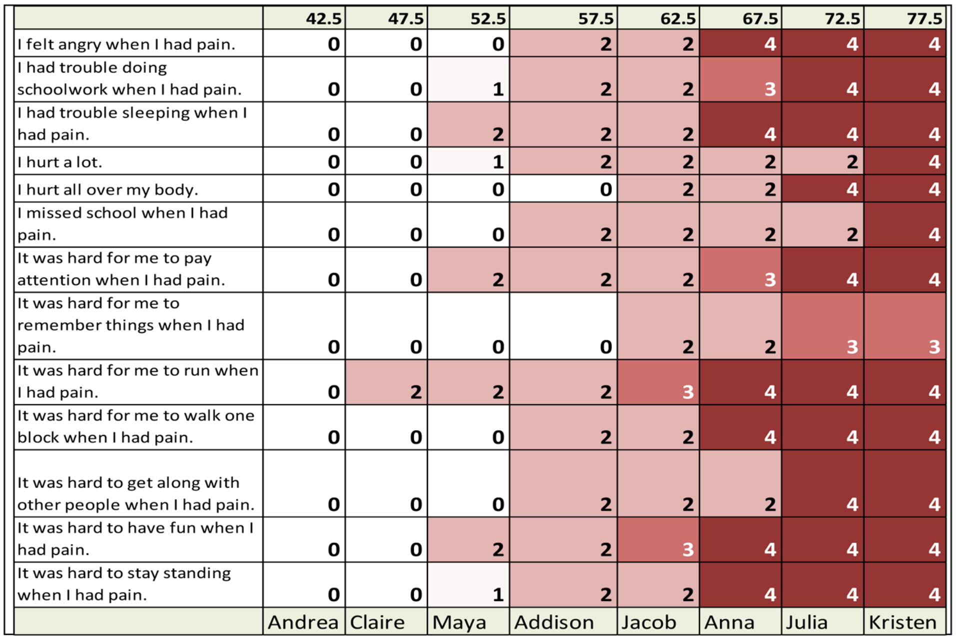
Item 1:	Never	Never	Rarely	Some-times	Some-times	Always	Always	Always
Item 2:	Never	Rarely	Some-times	Some-times	Always	Always	Always	Always
Item 3:	Never	Rarely	Rarely	Rarely	Rarely	Some-times	Some-times	Always
Complete for all items in bank								

Item 11	Item 5	Item 8	Item 1	Item 2	Item 5	Item 3	Item 4
Item 17	Item 7	Item 9	Item 6	Item 16	Item 7	Item 5	Item 7
Item 20	Item 13	Item 23	Item 18	Item 17	Item 23	Item 10	Item 18
Item 22	Item 19	Item 39	Item 30	Item 38	Item 35	Item 19	Item 30
Item 30	Item 31	Item 40	Item 31	Item 39	Item 36	Item 21	Item 39

In the last 7 days, Anne never got tired easily and never had trouble starting things because she was too tired. She felt that she was never so tired that it was hard for her to focus on work. She never found it hard to get out of bed in the morning because she was too tired and she never felt too tired to do sports or exercise.

Figure 1.

Steps for developing clinical vignettes from an IRT-calibrated item bank.



	42.5	47.5	52.5	57.5	62.5	67.5	72.5	77.5
I felt angry when I had pain.	0	0	0	2	2	4	4	4
I had trouble doing schoolwork when I had pain.	0	0	1	2	2	3	4	4
I had trouble sleeping when I had pain.	0	0	2	2	2	4	4	4
I hurt a lot.	0	0	1	2	2	2	2	4
I hurt all over my body.	0	0	0	0	2	2	4	4
I missed school when I had pain.	0	0	0	2	2	2	2	4
It was hard for me to pay attention when I had pain.	0	0	2	2	2	3	4	4
It was hard for me to remember things when I had pain.	0	0	0	0	2	2	3	3
It was hard for me to run when I had pain.	0	2	2	2	3	4	4	4
It was hard for me to walk one block when I had pain.	0	0	0	2	2	4	4	4
It was hard to get along with other people when I had pain.	0	0	0	2	2	2	4	4
It was hard to have fun when I had pain.	0	0	2	2	3	4	4	4
It was hard to stay standing when I had pain.	0	0	1	2	2	4	4	4
	Andrea	Claire	Maya	Addison	Jacob	Anna	Julia	Kristen

Figure 2.

Response probability table of PROMIS pediatric pain interference items for vignette creation. Response options: 0 = never, 1 = almost never, 2 = sometimes, 3 = often, 4 = almost always

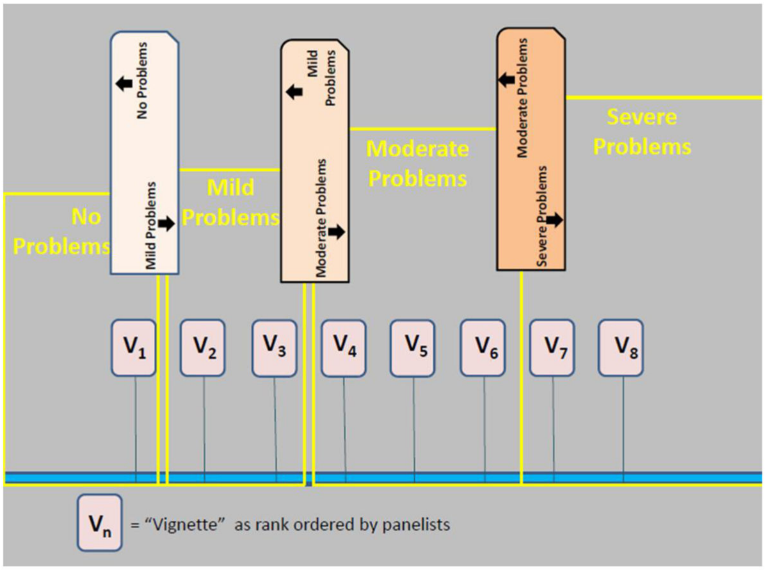
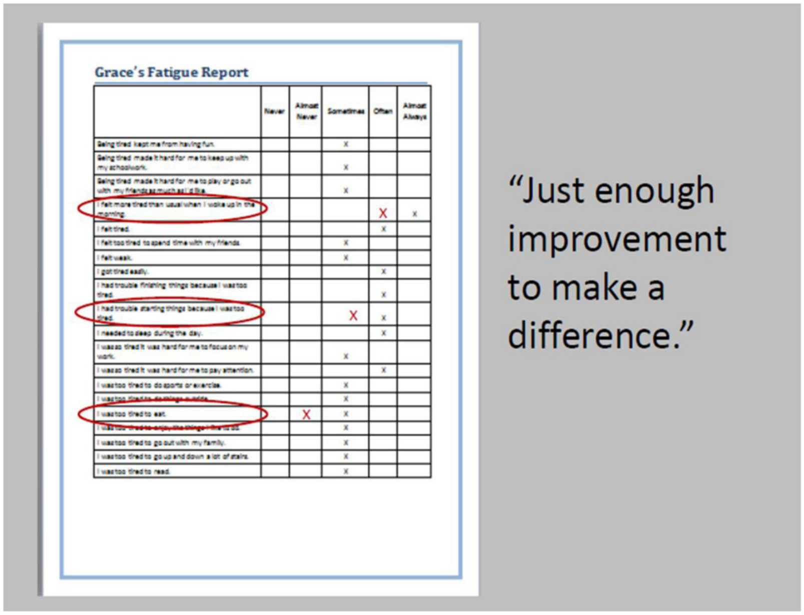


Figure 3.

Setting Cut Scores: After rank ordering vignettes by severity panelists established cut points between severity levels of health the problem.



Grace's Fatigue Report

	Never	Almost Never	Sometimes	Often	Almost Always
Being tired kept me from having fun.			X		
Being tired made it hard for me to keep up with my schoolwork.			X		
Being tired made it hard for me to play or go out with my friends as much as I'd like.			X		
I felt more tired than usual when I woke up in the morning.				X	X
I felt tired.				X	
I felt too tired to spend time with my friends.			X		
I felt weak.			X		
I got tired easily.				X	
I had trouble finishing things because I was too tired.				X	
I had trouble starting things because I was too tired.			X	X	
I needed to sleep during the day.				X	
I was so tired it was hard for me to focus on my work.			X		
I was so tired it was hard for me to pay attention.				X	
I was too tired to do sports or exercise.			X		
I was too tired to do things outside.			X		
I was too tired to eat.		X	X		
I was too tired to enjoy the things I like to do.			X		
I was too tired to go out with my family.			X		
I was too tired to go up and down a lot of stairs.			X		
I was too tired to read.			X		

"Just enough improvement to make a difference."

Figure 4.

Establishing Minimal Important Difference: Example scoring of "just enough to make a difference" in fatigue domain.

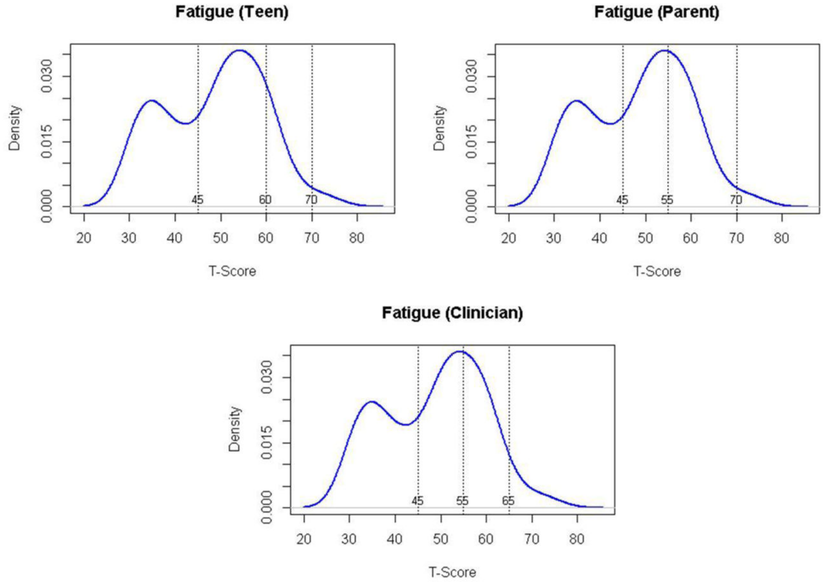


Figure 5a.

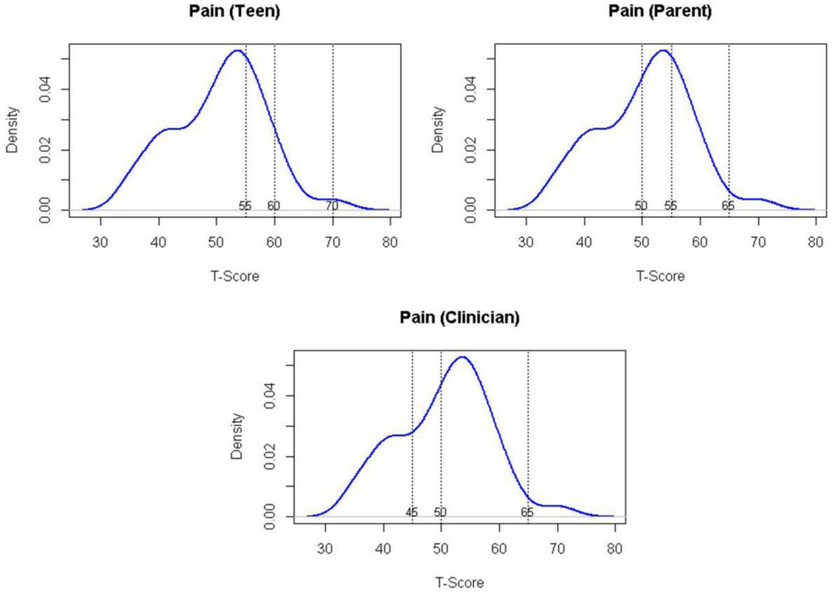


Figure 5b.

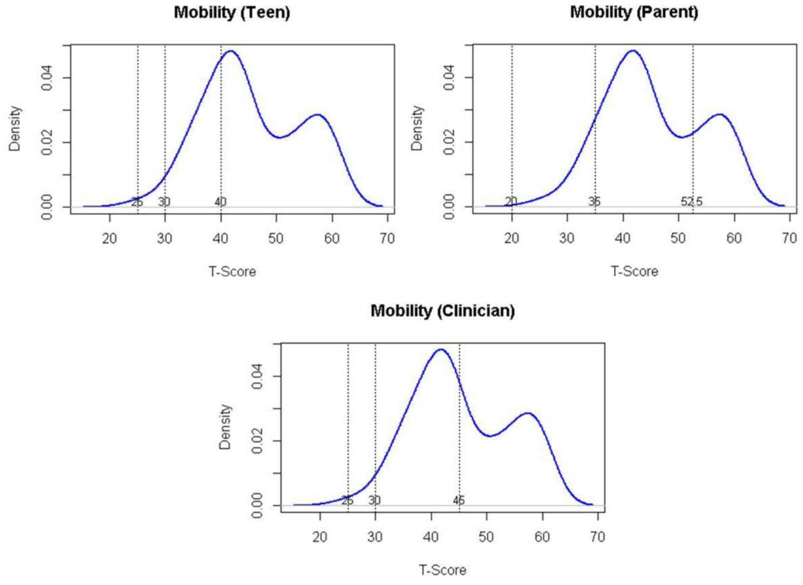


Figure 5c.

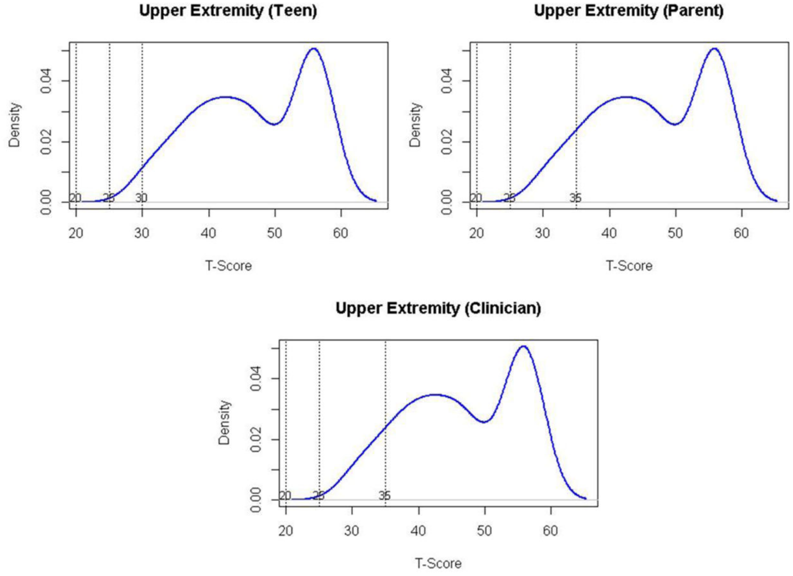


Figure 5d.

Figure 5.

Figure 5a. Distribution of Fatigue scores (y axis) by T score (x axis) for pwJIA panelists (top left panel), parent panelists (top right panel), and clinician panelists (bottom panel) with vertical lines separating severity categories (no problems, mild problems, moderate problems, severe problems).

Figure 5b. Distribution of Pain Interference scores (y axis) by T score (x axis) for pwJIA panelists (top left panel), parent panelists (top right panel), and clinician panelists (bottom panel) with vertical lines separating severity categories (no problems, mild problems, moderate problems, severe problems).

Figure 5c. Distribution of Mobility scores (y axis) by T score (x axis) for pwJIA panelists (top left panel), parent panelists (top right panel), and clinician panelists (bottom panel) with vertical lines separating severity categories (no problems, mild problems, moderate problems, severe problems).

Figure 5d. Distribution of Upper Extremity scores (y axis) by T score (x axis) for pwJIA panelists (top left panel), parent panelists (top right panel), and clinician panelists (bottom panel) with vertical lines separating severity categories (no problems, mild problems, moderate problems, severe problems).

Table 1

Panelist Demographics

	Patients	**Parents**	**Clinicians**
N =	4	5	7
% female	100%	100%	57%
age range	15–20	13–20 *	
years experience			5–35 years

PROMIS Mobility	M = 44.93 (sd = 12.9)
	Range = 32.8–58.5		

PROMIS Upper Extremity Function	M = 47.47 (sd = 3.4)
	Range = 43.6–50		

PROMIS Pain Interference	M = 45.7 (sd = 11.1)
	Range = 34–56		

PROMIS Fatigue	M = 43.23 (sd = 1.7)
	Range = 41.3–44.5		

*of their kids with JIA

Table 2

Demographics of Patient Data of Clinical Reference Population Sample, *N* = 121

	Mean (*SD*)
Age in years	13.0 (2.7)
Female, n (percent)	86 (71.1%)
JIA category, n (percent)	
Polyarticular	77 (63.6%)
Oligoarticular	21 (17.4%)
Systemic	11 (9.1%)
Enthesitis Related Arthritis	4 (3.3%)
Psoriatic	6 (5.0%)
Pediatric PROMIS T-Scores	
Pain Interference	51.1 (8.8)
Fatigue	47.5 (12.5)
Upper Extremity Function	45.7 (9.6)
Mobility	44.5 (9.8)

Table 3

Consensus cut-scores by domain and by expert panel.

Pediatric PROMIS Domain	Adjacent Categories	Ranges of Scores within Each Category			Percentages of Clinical Sample by Severity Level		
		pwJIA classifications	Parents classifications	Clinicians classifications	pwJIA classifications (%)	Parents classifications (%)	Clinicians classifications (%)
Fatigue	No problems	<45	< 45	< 45	47	47	47
	Mild problems	45–60	45–55	45–55	39	28	28
	Moderate problems	61–70	56–70	56–65	11	22	21
	Severe problems	>70	>70	>65	3	3	4
Pain Interference	No problems	<55	<50	<45	69	54	36
	Mild problems	55–60	50–55	45–50	19	15	18
	Moderate problems	61–70	56–65	51–65	11	27	42
	Severe problems	>70	>65	>65	1	4	4
Mobility	No problems	>40	>52.5	>45	68	29	53
	Mild problems	40–30	52.5–35	45–30	30	58	45
	Moderate problems	29–25	34–20	29–25	2	13	2
	Severe problems	<25	<20	<25	0	0	0
Upper Extremity Function	No problems	>30	>35	>35	95	87	87
	Mild problems	30–25	35–25	35–25	5	13	13
	Moderate problems	24–20	24–20	24–20	0	0	0
	Severe problems	<20	<20	<20	0	0	0

Table 4

Minimally Important Differences by domain, severity, and expert panel.

Pediatric PROMIS Domain	T-Score	Name	Severity Classification	Mean MID (SD)		
				pwJIA	Parents	Clinicians
Fatigue	77.5	Owen	severe	5.43 (3.32)	9.42 (3.24)	3 (3.24)
	67.5	Grace	moderate	3.65 (3.56)	3.5 (3.44)	1.37 (3.56)
	57.5	Jessica	mild	4 (3.44)	4.8 (3.49)	2.97 (3.54)
Pain Interference	72.5	Julia	severe	7.55 (4.27)	12.68 (4.09)	5.33 (4.41)
	67.5	Anne	moderate-severe	5.8 (3.73)	8.08 (3.71)	3.35 (3.77)
	57.5	Addison	mild-moderate	3.3 (3.32)	5.48 (3.13)	2.07 (3.37)
Mobility	12.5	Kylie	severe	3.73 (2.41)	4.37 (2.39)	2.18 (2.55)
	27.5	Emily	moderate	0.1 (1.20)	1.27 (1.60)	0.08 (1.22)
	37.5	Olivia	mild	5.03 (3.79)	5.4 (3.91)	1.6 (3.41)
Upper Extremity Function	17.5	Chloe	severe	2.58 (2.60)	2.83 (2.81)	1.83 (2.61)
	22.5	Brianna	moderate	0.05 (1.21)	2.17 (1.78)	0.35 (1.52)
	37.5	Emma	none-mild	3.08 (2.92)	3.57 (3.12)	1.47 (2.65)