Best Practice & Research Clinical Rheumatology 33 (2019) 101436

Contents lists available at ScienceDirect

Best Practice & Research Clinical Rheumatology

journal homepage: www.elsevierhealth.com/berh

7

Taking the patient and the patient's perspective into account to improve outcomes of care of patients with musculoskeletal diseases

Martijn A.H. Oude Voshaar [a, b, *], Mart A.F.J. van de Laar [a, b]

[a] Department of Psychology, Health & Technology, University of Twente, the Netherlands
[b] Transparency in Healthcare, University of Twente, Hengelo, the Netherlands

Keywords:
Item response theory
Patient reported outcomes
Item banking
Validity
Reliability
Outcomes

A B S T R A C T

Patient-reported outcome measures are commonly used in the assessment of patients with musculoskeletal diseases. The present review provides an overview of historic and recent developments, including core set recommendations for assessing patient-reported outcomes in patients with fibromyalgia, osteoarthritis, rheumatoid arthritis, ankylosing spondylitis, and psoriatic arthritis. The evidence supporting commonly used patient-reported outcomes measures is reviewed. Furthermore, various methodological approaches that can be utilized to evaluate validity and measurement precision of patient reported outcomes are introduced. Commonly used methods based on the classical test theory as well as modern approaches based on item response theory will be discussed. The review finally describes the increasing use of item response theory-based approaches used in patient-reported outcomes assessment in the musculoskeletal diseases.

* Corresponding author. Transparency in Healthcare, University of Twente, Hengelo, the Netherlands.
E-mail address: a.h.oudevoshaar@utwente.nl (M.A.H. Oude Voshaar).

https://doi.org/10.1016/j.berh.2019.101436

What are patient-reported outcomes measures and for which purposes are they used in patients with musculoskeletal diseases?

Musculoskeletal diseases (MSK) are major contributors to the global burden of disease [1]. The present review summarizes the literature on patient-reported outcomes in MSK with a particular focus on (polyarticular) osteoarthritis, fibromyalgia, and various inflammatory arthritides (IA) (i.e., rheumatoid arthritis (RA), ankylosing spondylitis (AS), and psoriatic arthritis (PsA)) which are among the most studied MSK in recent times. Outcomes of interest in these diseases, including pain, fatigue or experienced difficulties performing daily activities, cannot be directly observed as they are only known to the patient. Standardized assessments of subjective outcomes based on responses provided directly by patients themselves, usually in the form of questionnaires, without subsequent interpretation or alteration of the responses by anyone else are known as patient-reported outcomes measures (PROMs) [2]. For an increasing number of outcome domains in MSK, PROMs are the accepted standard and a large literature exists that supports the validity and reliability of PROMs applied in these populations [3]. Even for outcomes domains for which 'objective' or physician-reported measures exist as well, PROMs offer a number of advantages [4]. For example, PROMs usually come with lower costs, less professional time, and no training is typically required for them to be implemented. Other advantages of PROMs over physician-reported outcomes are that PROMs are not affected by observer bias, which is of particular interest in for example quality assurance initiatives where health professionals might assess patients from their own practice [1]. Moreover, patients only need to complete their own PROM, whereas health professionals would have to complete physician-reported measures for all their patients.

Large-scale assessments using PROMs were first introduced in the field of health services research as a means to meet increasing societal demand for insight into the effects of healthcare on the lives of patients. This demand resulted from increased healthcare spending in the second half of the 20th century [5]. A variety of generic and disease-specific PROMs were developed during that time to inform resource allocation decisions and to help healthcare providers and executives demonstrate value created for their patients in terms of outcomes of interest to a wider audience than physicians alone [6]. Because of their widespread use, PROMs also became popular in clinical trials, and are now increasingly used by pharmaceutical companies to support labeling claims for newly developed medical products [7]. In recent times, the International Consortium for Health Outcomes Measurement (ICHOM) has been working to implement PROMs in daily clinical practice to facilitate sharing of outcomes data between healthcare providers for shared learning and global benchmarking [8].

Which domains of patient-reported outcomes can be distinguished and which of them are important in patients with MSK?

PROMs can be used to measure the impact of health status on quality of life, i.e. to assess Health Related Quality of life (HRQOL) [9]. While no comprehensive definition of HRQOL presently exists, a useful structured representation of the HRQOL spectrum, as well as the ways in which different HRQOL domains might interact is provided by the International Classification of Functioning, Disability and Health (ICF) [10]. The ICF framework considers health from a biopsychosocial perspective, in accordance with the WHO definition of health as a "state of complete physical, mental and social well-being and not merely the absence of disease and infirmity". The ICF is a taxonomy with Body functions, Activities, and Participation as three main components, which are hierarchically related concepts that refer to different levels of consequence of disease, ranging from the biological to the individual and the societal perspectives. Each component contains a number of subcomponents, referred to as chapters, which can be further differentiated into codes. Each code refers to a specific health concept.

The body functions component pertains to the biological perspective and encompasses all physiological or psychological functions of body systems. Loss or abnormalities of body functions are referred to as impairments (i.e. signs and symptoms of disease). Subjectively experienced impairments (i.e. symptoms) such a pain and fatigue are common PROM domains since the patients themselves only know information on these outcomes. The second ICF component "Activities" pertains to the individual level and refers to the execution of specific everyday tasks by an individual. This component is

concerned with activities a person may engage in during their daily lives, including self-care, mobility and domestic life (household, work and chores). Activity limitations are commonly assessed by measures of physical functioning in MSK. The final component 'participation' refers to engagement of individuals in social roles. Participation restrictions occur when disease-related impairments or activity limitations prevent an individual from participating in a social role. In the ICF framework participation restrictions may involve major life areas (education or work), community, social or civic life (e.g. engaging in leisure activities) and interpersonal interactions and relationships (e.g. maintaining family relationships). Several PROMs exist to assess participation restrictions in social roles [11] or work disability [12].

As made clear by the foregoing, many different patient-reported outcome domains can be distinguished that could be assessed in studies of patients with MSK. To avoid difficulties in aggregating results from different studies due to inconsistent outcome reporting, the most relevant outcome domains for a given research purpose may be summarized in international, consensus based core set recommendations [13]. For various MSK conditions, core sets of outcome domains to be assessed in clinical trials have been developed [14,15]. Van Tuyl et al. summarized PRO domains that are included in these core sets [16]. She found that pain, physical function, and patient global assessment were outcomes that were included in the clinical trial core sets for RA, ankylosing spondylitis, psoriatic arthritis osteoarthritis and fibromyalgia. Fatigue is included in the RA and FM core sets. According to a 2009 Systematic review, the majority of RA clinical trials do indeed report on physical function (89%), patient global assessment (61%) and pain (56%), while fatigue was not yet frequently reported [17]. In a more recent literature review of 250 articles by Kilic et al., in which also papers reporting on studies other than clinical trials ($n = 137$) were included, similar results were found, with physical function again the most common PROM, followed by patient global assessment and pain, while other domains including fatigue were less frequently assessed [18]. In 2017, EULAR published a core data set recommendation for observational studies which included the PRO domains pain, PGA, physical function and quality of life [19].

Recommendations for use of PROMs in daily clinical practice of patients of patients with IA were recently provided by the ICHOM Inflammatory arthritis working group [20]. When compared with clinical trial core sets, the ICHOM Standard Set for IA contains more PROMs. This is likely in part due to the inclusion of patients in the working group as well as ICHOM overall focus on outcomes that matter to patients. The ICHOM working group recommended assessment of pain, fatigue, physical function, overall physical and mental health impact, and work/school/housework ability and productivity [20].

Across the various core sets, pain, fatigue, and physical function are the most common PROM domains, besides the overall assessment. These outcomes are also most frequently regarded as important by patients according to various qualitative studies in which patients are asked to list or rank the most important aspects of their disease [20–22]. Because of their importance across different MSK, the remainder of this review will focus on the PROM domains pain, physical function and fatigue.

Which quality criteria are important for PROM measures to possess?

Validity

An assessment of the validity of PROM scores should provide information on the degree to which the PROM scores are useful to answer the question of interest. While different methods have been proposed to assess validity, these should not be considered as different forms of validity, but rather as different ways to scientifically support particular interpretations of PROM scores [23]. Content validity is one broad group of validation techniques that refers to an examination by experts of the degree to which the PROM items are relevant to and representative of the domain the PROM intends to measure [10]. According to the International society for quality of life research (ISOQOL), content validity of a PROM should be supported by evidence that patients and/or experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application [24]. The second important type of validity evidence is construct validity. Construct validity

has different aspects, each directed at particular conditions that need to be met in order for PROM scores to yield useful information. When a new PROM is developed or when a PROM is first introduced in a new patient population, support should be provided for the underlying measurement model of the PROM. Most PROMs are multiitem questionnaires that have one or several (sub)scales that (all) intend(s) to measure a single (sub)domain of interest. Exploratory, bifactor and confirmatory factor analysis, as well item response theory IRT based techniques such as the Rasch measurement approach and the non-parametric IRT models introduced by Mokken can be used to test or check whether response data collected from the target population correspond with the intended measurement model. These techniques also provide information on the degree to which the total score provides a useful summary of the (sub)domain measured by the (sub)scale and can be used to weed out weak items. Although these procedures are complex, there are various resources available to help researchers perform and interpret Mokken scaling analysis [25,26], Rasch analysis [25] exploratory factor analysis [27,28], and bifactor analysis [29,30].

Once the structure of a PROM is well supported, a next step in validation research is to provide evidence that supports intended interpretations of the PROM scores. The intended purpose of many PROMs is to document changes in status over time. The validity of change scores (responsiveness) is therefore frequently particularly important when evaluating PROM score validity evidence. Responsiveness can be supported by showing that the magnitude of PROM change scores that occur over the course of an intervention correspond with a priori expectations of the magnitude of change in the construct being measured [4]. According to ISOQOL, the responsiveness of a PROM should at minimum be supported by evidence that changes in scores are consistent with predefined hypotheses regarding changes in the measured PROM in the target population for the research application [24]. Similarly, evidence should be gathered about the degree to which the correlations of PROM scores with other measures are consistent with those expected. Construct validation is an iterative process in which confidence in the degree to which a PROM actually reflects the construct it intends to measure increases as applications of the measure consistently yield results that would be expected, given theories about how the construct of the PROM being considered relates to other constructs [40]. Especially for newly introduced PROMs, proper evaluation of construct validity requires investigators to be specific about expected relations among instruments included in the validity assessment; taking into account that the relations between the domains being measured, measurement error and method of measurement all contribute to the observed correlations. Ideally, explicit hypotheses regarding the magnitude of the expected relations (e.g. correlations) between the instrument to be validated and the instruments used for validation should be defined before data is collected [31].

Measurement precision

Besides systematic biases, observed PROM scores will be affected by random measurement errors as well and therefore the PROM score an individual patient achieves should be considered a more or less accurate estimate of their true score. Classical test theory (CTT) is a psychometric framework that can help understand the degree to which total PROM scores are affected by measurement errors (see box 1). If repeated measurements are available of the same group of patients, the test-retest correlation coefficient can be used to estimate reliability of the scores. The time interval between assessments, together with the magnitude of the resulting test-retest (intraclass) correlation coefficient provide an indication of the degree to which scores can be generalized over time [11]. If only one measurement is available and the PROM has multiple items, a lower bound estimate of the reliability coefficient can be calculated from the inter-item covariance matrix. Such reliability coefficients assume that the individual items making up the PROM are essentially true score equivalent, which means that true scores on the items are perfectly correlated but allowed to differ by a constant [32]. The most popular coefficient, Cronbach's alpha, is known to underestimate this lower bound to a larger degree than less commonly used alternatives, including the greatest lower bound reliability coefficient and coefficient omega [33,34]. It has been recommended that for research purposes, reliability coefficients should generally exceed 0.80 [35].

It should be noted that reliability coefficients pertain to a group of patients and should therefore not be used when the precision of individual patient scores is of interest. Instead, for individual patient

Box 1

CTT states that a patient's (denoted *i*) observed PROM score (O_i) can be decomposed into the patient's "true" score (T_i) on the instrument and a random measurement error component (M_i), i.e. $O_i = T_i + M_i$. Because M_i is defined to be random, the expected value (mean) of a distribution of measurement error scores over (hypothetical) repeated administrations of the same PROM to patient $i = E(M_i) = 0$, hence $E(O_i) = T_i$, i.e. the observed score is an unbiased estimate of the true score. Furthermore, it is assumed that error scores and true scores are uncorrelated. This assumption allows the observed score variance in a group of patients to be decomposed into a true score component and a measurement error component: $\sigma_O^2 = \sigma_T^2 + \sigma_M^2$. The overall precision of the scores in a population of patients can be quantified using reliability coefficients that represent the proportion of the observed score variance that is accounted for by variation among patients in the attribute of interest, i.e.: $r = \frac{\sigma_T^2}{\sigma_O^2}, = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_M^2}$. However, because neither σ_T^2 $nor \sigma_M^2$ can be directly observed, these quantities have to be estimated themselves before the reliability coefficient can be calculated.

scores, 95% confidence intervals can be calculated by multiplying the standard error of measurement with 1.96 [36]. The SEM is defined as: $SEM = \sigma_O \sqrt{(1 - r)}$.

Although most psychometric studies in MSK research to date have relied on CTT for the assessment of measurement precision, the CTT framework has well-known limitations. CTT based reliability coefficients are not only influenced by characteristics of the PROM but to a large extent by the heterogeneity of the group of patients used to estimate them as well. Consequently, the generalizability of CTT based estimations of precision is limited. Moreover, the standard error of measurement of individual scores are based on the unrealistic assumption of equal error scores for individual patients.

These limitations are addressed in the modern test theory framework [37]. Within this item response theory (IRT) framework, responses probabilities for each of an item's response options are predicted using nonlinear functions of the latent variable. Scores can be estimated using maximum likelihood or Bayesian procedures. The uncertainty about an individual' score is reflected in the standard error of the maximum likelihood estimate or the standard deviation of the posterior distribution when Bayesian scoring procedures are used. Because of the non-linear nature of IRT models, these precision estimates vary as a function of the latent variable. Consequently, confidence intervals for individual score interpretations (which are calculated from the standard error of estimation) are no longer of the same length for all patients. Instead, the length of the confidence intervals depends on the latent trait value of the patient and the characteristics (item parameters) of the items used to estimate the score. The precision of an IRT scored PROM can therefore best be described over the different score levels that be distinguished in a graph or a table with the standard errors or posterior standard deviations for different score levels [38]. However, more commonly, the information functions are shown. Information functions are defined as the expected value of the inverse of the error variance of the maximum likelihood estimate. Each item has its own item information function which is calculated from the item parameters. An item information function describes how much information an item contributes to the IRT score, across the latent variable scale. The main advantage of information over the standard error is that different item information functions can be added together to form the score information function. For most two parameter IRT model-based scales, an information value of 10 corresponds to a standard error of 0.32, which in turn corresponds to 90% true score variance or a reliability coefficient of 0.90 [38].

Feasibility

Researchers intending to use a PROM for a particular purpose should first assess whether the required resources are available for the PROM to be implemented in that setting. Most PROMs can be administered electronically or by hand. In individual studies, self-administered and computerized versions of the same PROM are sometimes used interchangeably. In addition, patients are sometimes

assisted by health professionals when filling out PROMs. Various studies that have examined the between method of administration variability of scores were meta-analyzed recently by Rutherfeld et al. They found that mode of administration does not cause bias in patient-reported outcome results [39]. Rutherfeld et al. considered the mean differences in PROM total scores. Several studies were also performed in which more detailed, item level IRT based analysis of differential item functioning was performed [40]. Together, these studies suggest that mixed modes of administration do not lead to biased results and different modes of administration might be combined within one study. This might sometimes increase the number of patients that can be recruited, especially in resource scarce settings. Electronic PROMs are almost by definition automatically scored, while self-administered PROMs sometimes require complex scoring procedures, including those based on IRT that required specialized software or access to (copyrighted) item calibrations and CAT algorithms. In addition, there is an increasing number of PROMs that require a license fee for particular applications [20]. Comparability of outcomes with previous studies is also a factor to consider when selecting PROMs for a new study.

There are also several aspects of patient burden that need to be considered. According to the ISOQOL recommendations, patients should not be subjected to overly long questionnaires or to overly frequent data collections [24]. The emotional impact filling out PROMs may have on patients should also be considered. It has been shown that distressful survey content can negatively affect mood and increase stress in patients, particularly those with pre-existing emotional vulnerabilities [41]. It is also important to consider the literacy levels of the target population. In general, it is believed that PROM items should be written at 6th grade education level (i.e. suitable for children 11–12 years) [24]. The grade level of a PROM text could be estimated using the Flesch-Kincaid grade level estimate [42] that is embedded in the Microsoft Word, word processing software. It is also recommend to pretest PROMs when they are newly developed or first introduced in a new population. Two major types of cognitive interviewing methods are available that can be used to verify that all items are understood as intended by patients; think-aloud interviewing and verbal probing techniques [43,44]. In the verbal probing approach, structured interviewing techniques are used to focus the attention on areas considered of interest by an interviewer, allowing the interviewer to focus on particular areas that appear to be relevant as potential sources of response error. In the think aloud approach, patients are encouraged to work through the PROM while verbalizing their thoughts. The main advantage of think-aloud methods is that there is minimal interviewer-imposed bias, and, consequently, unanticipated problems in the response behavior of participants are more likely to be detected. These two methods therefore might complement each other: Verbal probing techniques provide a good way to verify that questions are comprehended as intended on a semantic and conceptual level in a structured setting where the interviewer controls the course of the exchange. On the other hand, think aloud methods are useful to learn about unanticipated problems [45].

What are suitable measures for important outcomes in RA?

Pain

Although it is widely recognized that pain is a multidimensional phenomenon, pain is almost exclusively measured using unidimensional tools (i.e. pain intensity) in clinical trials in the various IA, as well as OA, while for FM more elaborate, multidimensional pain PROMs have additionally been applied [20,46]. In OA trials, the WOMAC pain scale has sometimes been used to measure pain intensity. However, single item pain scales such as the numerical rating scale (NRS) or visual analogue scale (VAS) pain intensity are almost exclusively used in the other MSK reviewed here. Single item PROMs have the advantage that they are fast to administer and easy to score. Moreover they are usually highly sensitive to change [47] and yield reliable scores (i.e. test-retest reliabilities usually $r > 0.7$) [20]. The NRS is usually the preferred single item pain scale because it is easy to administer verbally or by phone in contrast to the VAS [20,48]. A concern with single items NRS and VAS measures is that various different formulations for the item and item response options are used, which makes it difficult to compare results between studies. The ICHOM IA group has proposed a unified formulation for the NRS

pain, to be used in patients with IA [20]. Ten Klooster et al. showed that patients whose VAS pain scores improve by at least 55% tend to view this improvement as satisfactory [49,50].

More comprehensive pain assessment measures such as the Multidimensional Pain inventory (MPI) and brief pain inventory (BPI) are also are also available [48]. More recently, the PROMIS pain interference, pain behavior item banks have become available. Although individual PROMIS item banks are unidimensional, these pain related item banks assess constructs other than pain intensity, which could be useful to provide more in depth characterization of patients' pain experience. However with some exceptions, their measurement properties have not been studied in much detail in the various MSK populations considered in this review [48,51–53]. The brief pain inventory has been shown to yield valid and reliable scores in osteoarthritis [54,55] and was found to be more responsive to change compared with the PROMIS pain interference scales [56]. Preliminary evidence about the PROMIS pain interference item bank is also available for RA [54].

In addition to intensity and impact of pain on daily lives of patients, it is sometimes of interest to assess the type of pain experienced by patients. While objective measures, including quantitative sensory testing are usually considered the gold standard, several PROMs are also available that can be used to given characterize the presence of particular types of pain or to monitor outcomes in terms of these types of pain. PAINdetect is a PROM that can be used to identify neuropathic pain components [57]. More recently, the Generalized Pain Hypersensitivity questionnaire has been developed to identify patients with possible central augmentation of pain [10].

Physical function

Many different physical function PROMs are available with documented evidence regarding their measurement properties in the MSK reviewed here [3,51,52,58,59]. In RA clinical trials, various versions of the Health Assessment Questionnaire (HAQ) disability Index are commonly used that are all free to use and can be administered in less than 5 min [60]. The validity and reliability of HAQ-DI, MDHAQ and HAQ-II scores in patients with RA are supported by a large body of evidence, and the items are easy to understand for patients with different literacy levels [20,61]. The original HAQ-DI has >60 translated versions. However, most of the items of the original HAQ refer to basic activities of daily living. Hence, ceiling effects are a problem, particularly so in patients with controlled disease [3,62]. It is therefore recommended to use the multidimensional HAQ (i.e. MDHAQ) or HAQ-II).

For patients with ankylosing spondylitis the BASFI is probably the most commonly used physical function instrument. However, the Dougados Functional Index is an alternative option with items that are likely to be easier to read. Measurement properties of both instruments were found to be well supported by a systematic review [52]. BASFI and various HAQ versions are also commonly used in psoriatic arthritis trials [20]. Various physical function PROMs, including the pain interference scales of the BPI, the Sheehan disability scale and the FIQ physical function scale have been recommended for fibromyalgia [58]. However, none of these scales were reviewed in the only available systematic review from 1990. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) has been among the most commonly used physical function PROMs in OA clinical trials. Various studies are available that have shown WOMAC to yield reliable scores and that is has favorable psychometric properties in OA populations [63]. Disadvantages of WOMAC are that licensing fees apply and that different versions, including with different response options exist. Limited information is also available in the public record regarding these different versions, which makes it challenging to compare outcomes between studies. The lower extremities functional scale has been recommended as an alternative physical function PROM in OA [60,64].

In principle, generic physical function measures such as the SF-36 and PROMIS physical function are also available but are rarely used as endpoints in clinical studies in these populations, possibly because their item content does not match sufficiently well with areas of particular clinical significance for any of the patient populations. For instance HAQ-DI is widely considered to have high content validity for patients with RA because it has various items relating to dexterity, whereas BASFI focuses on activities relating to the functioning of the central region of the body, which is more relevant for patients with AS. The English version of all the physical function PROMs reviewed in this section can be downloaded from the EULAR outcome measures library, free of charge.

Fatigue

Like pain, fatigue is frequently assessed with single item, fatigue severity measures such as a VAS or NRS. Although several studies support the validity and reliability of single item fatigue instruments [65], it has also been shown that such measures yield less precise scores for patients with fatigue levels on the lower and higher ends of the fatigue spectrum in patients with inflammatory arthritis [66]. Based on IRT analysis, it was recommended that the SF-36 vitality scale could be used for patients with mild to moderate fatigue, while the multi-item BRAF-MD and FACIT were found to be well suited for patients with moderate to high fatigue. Some authors have argued that fatigue is a multidimensional construct. However, several studies show that item response data of multiple multi item fatigue instruments combined can still be described using a one-dimensional measurement model [67].

Future directions (IRT)

Most PROMs that are now in widespread use contain a fixed set of items that is intended to be administered to each patient. A key disadvantage of such measures is that they can only be used to compare patients who have responded to all the items. It is generally not possible to compare scores of groups of patients who have responded to different PROMs. Individual patient scores can only be compared over time or between patients if there are no missing individual responses, unless exactly the same responses are missing for all comparisons.

Many recent developments in PROM research rely on item response theory based approaches, where such problems do not exist. With IRT, large numbers of items that measure the same construct can be calibrated to a common scale. In this way, IRT can be used to develop universal metrics on which (potentially all known) items of a particular PROM domain can be mapped. This has several useful applications that are increasingly being explored in the field outcomes research.

Firstly, this allows outcomes to be compared between studies in a way that is otherwise only possible if the same PROM was used in all studies. In several early studies, the items of two PROMs were linked to a common scale, in order to produce a crosswalk table for converting scores [68,69]. More recently several larger projects have been completed in which the items of multiple PROMs were linked to a common scale. For instance, the items of 10 commonly used physical function PROMS, including multiple versions of HAQ, BASFI, Funktionsfragenbogen Hannover, SF-36 physical function scale, Rasch assessment of everyday activity limitations, PROMIS physical function and a NRS were calibrated to a common scale [70]. This project supported the ICHOM IA initiative in which a standard data set was developed to support widespread collection and reporting of outcomes data in clinical practice. The standardized physical function reporting metric allows healthcare providers and researchers to report outcomes obtained using any of the physical function measures endorsed by the ICHOM IA working group and still compare outcomes with other professional who have used a different instrument.

In the PROSETTA project several commonly used “legacy” (i.e. previously existing fixed length) measures of various PROM domains were linked to the corresponding PROMIS item banks. For instance the HAQ-DI and SF-36 physical functioning scale were linked to PROMIS physical function [71]. The goal of the PROSETTA project was to help researchers interested in using the PROMIS instruments to transition from legacy measures to PROMIS. Similar initiatives were performed for fatigue and pain [72].

There is also an ongoing project in which ICF-based common metrics are being developed to allow standardized documentation of functioning information in national health information systems. In that project, the content of various PROMs is linked to the ICF using a standardized approach [73]. Items that pertain to similar ICF chapters are then jointly calibrated to facilitate standardized chapter reporting for professionals that have used different instruments [74,75].

IRT is not only useful for standardization but also for developing targeted optimally precise measures. This is achieved by making use of the fact measurement precision is locally defined in IRT. Two different methods have been described in the literature that help researchers to develop psychometrically optimal measures by automating the item selection process. Computerized adaptive testing (CAT) is the best-known procedure in which scores of individual respondents are optimized in real-time by having a computer select items from an item bank that provide the

most information about a respondent's trait level, given their the estimated level of the respondent based on item responses up until that point [76,77]. If there are sufficient numbers of high-quality, well-matched items in the item bank, CAT allows shorter, yet more precise assessments of patient reported outcomes, compared with fixed length instruments. The Amsterdam Linear Disability score is an early example of a physical function item bank for which a CAT algorithm has been developed [78]. The PROMIS project is the first initiative aimed at bringing the advantages of IRT and CAT to health outcomes measurement. In that project, a domain framework was first developed that intends to comprehensively capture the various PROM outcome domains that are in widespread use. Subsequently, item banks and CAT were developed to assess individual domains. In principle, the PROMIS item banks are fee to use, although assessment using CAT may require access to an online platform for which a fee is still payable. Accumulating evidence suggests that assessment based on the PROMIS item banks can yield better results compared with routine measures in the MSK populations [79–91]. However, relatively few studies have yet examined the responsiveness of CAT versus short-forms. Besides PROMIS there is a multidimensional CAT for fatigue [92].

Optimal Test Assembly (OTA) is an alternative approach for deriving fixed-length questionnaires. It differs from CAT in that the most suitable items are selected for a range of trait levels that are considered relevant a priori [93]. The result of the process is a fixed-length instrument that is optimal with respect to the precision of scores for the defined population as well as a number of other constraints. OTA methods retain many of the benefits of CAT and can be applied in settings in which there is no access to the required infrastructure for CAT. Thus far, OTA methods have been in a single study in the reviewed conditions. In that study, it was shown that an OTA based measure of physical function outperformed the HAQ-DI in terms of ceiling and floor effects and measurement precision, with only 6 items. However, in that study, it was also shown that CAT performed better than both HAQ-DI and OTA [94].

Summary and conclusions

While traditionally PROMs featured predominantly in clinical trials, the increasing interest of various stakeholders in outcomes of MSK patients has resulted in their widespread use in longitudinal observational registries, comparative effectiveness research and daily clinical practice as well. Various Core/Standard Set recommendations described in the present review offer guidance on the most relevant patient reported outcome domains in various MSK populations. Typically, researchers must choose between several PROMs that are available for the outcome domain they wish to assess. To help researchers and clinicians select the best PROM for use in their research projects we described which measurement properties are important to consider when reading PROM validation studies and, whenever possible have, provided quality criteria for these measurement properties. We finally described that researchers are increasingly making use of IRT based methods to develop standardized item banks for different patient reported outcome domains. CAT, OTA and linking different PROMs to a common scale are several applications of IRT based item banking that will no doubt shape how PROMs are used in the future.

Practice points

- Patient-reported outcome measures are increasingly common in clinical trials, clinical practice and assessment of quality of care of patients with musculoskeletal diseases.
- Pain, fatigue, physical function, and patient global assessments important PROM domains that are relevant across the various musculoskeletal disorders.
- Selection of suitable Patient reported outcome measures should be based on documented evidence regarding their measurement precision and the validity of the scores.
- Patient reported outcomes research increasingly makes use of modern test theory and computerized adaptive testing.

Research agenda

- More head-to-head comparisons of responsiveness of CATs versus fixed length PROMs should be performed.
- Domain specific item banks that include items of commonly used PROMs for specific domains should be developed to facilitate score comparisons between studies and development of targeted instruments for different research populations.
- Existing PROMs can be further refined using Optimal Test Assembly based methodology.

Conflicts of interest

None of the authors have any conflict of interest in relation to this work.

Funding

No funding was received for this work.

References

[1] Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346. f167–f167. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23358487. [Accessed 18 January 2019].
[2] Speight J, Barendse SM. FDA guidance on patient reported outcomes. BMJ 2010;340. c2921–c2921. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20566597. [Accessed 4 January 2019].
[3] Voshaar M, Klooster PM ten, Taal E, et al. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes 2011;9. Available at: %3CGo.
[4] Wolfe F, Pincus T. Listening to the patient: a practical guide to self-report questionnaires in clinical care. Arthritis Rheum 1999;42:1797–808. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10513792. [Accessed 25 January 2019].
[5] Ellwood PM. Shattuck Lecture–outcomes management. A technology of patient experience. 1988. Arch Pathol Lab Med 1997;121:1137–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3367968. [Accessed 4 January 2019].
[6] McHorney CA. Health status assessment methods for adults: past accomplishments and future challenges. Annu Rev Public Health 1999;20:309–35. Available at: http://www.annualreviews.org/doi/10.1146/annurev.publhealth.20.1.309. [Accessed 4 January 2019].
[7] Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Contr Clin Trials 2004;25:535–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15588741. [Accessed 4 January 2019].
[8] Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016;374:10–2. Available at: http://www.nejm.org/doi/full/10.1056/NEJMp1511701?query=featured_home.
[9] Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995;273:59–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7996652. [Accessed 29 December 2017].
[10] World Health Organization. International classification of functioning, disability and health world health organization geneva ICF ii WHO library cataloguing-in-publication data international classification of functioning, disability and health. Geneva: ICF; 2001. Available at: http://apps.who.int/iris/bitstream/handle/10665/42407/9241545429.pdf. [Accessed 23 January 2019].
[11] Oude Voshaar M, van Onna M, van Genderen S, et al. Development and validation of a short form of the social role participation questionnaire in patients with ankylosing spondylitis. J Rheumatol 2016;43:1386–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27182067. [Accessed 24 January 2019].
[12] Reilly M, Zbrozek A, Dukes E. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993;4:353–65.
[13] Boers M, Kirwan JR, Wells G, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24582946. [Accessed 8 January 2019].
[14] Boers M, Tugwell P, Felson DT, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol Suppl 1994;41:86–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7799394. [Accessed 17 May 2015].
[15] Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8507213. [Accessed 27 May 2015].
[16] van Tuyl LHD, Boers M. Patient-reported outcomes in core domain sets for rheumatic diseases. Nat Rev Rheumatol 2015;11:705–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26324860. [Accessed 28 December 2016].

[17] Kalyoncu U, Dougados M, Daures J-P, et al. Reporting of patient-reported outcomes in recent trials in rheumatoid arthritis: a systematic literature review. Ann Rheum Dis 2009;68:183–90. Available at: http://ard.bmj.com/cgi/doi/10.1136/ard.2007.084848. [Accessed 1 September 2016].
[18] Kilic L, Erden A, Bingham CO, et al. The reporting of patient-reported outcomes in studies of patients with rheumatoid arthritis: a systematic review of 250 articles. J Rheumatol 2016;43:1300–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27084908. [Accessed 8 January 2019].
[19] Radner H, Chatzidionysiou K, Nikiphorou E, et al. 2017 EULAR recommendations for a core data set to support observational research and clinical care in rheumatoid arthritis. Ann Rheum Dis 2018;77:476–9. Available at: http://ard.bmj.com/lookup/doi/10.1136/annrheumdis-2017-212256. [Accessed 8 January 2019].
[20] Oude Voshaar MAH, Gupta ZD, Bijlsma JWJ, et al. The international Consortium for health outcome measurement (ICHOM) set of outcomes that matter to people living with inflammatory arthritis consensus from an international working group. Arthritis Care Res (Hoboken) 2018. Epub ahead of print. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30358135. [Accessed 8 January 2019].
[21] Klooster PM ten, Veehof MM, Taal E, et al. Changes in priorities for improvement in patients with rheumatoid arthritis during 1 year of anti-tumour necrosis factor treatment. Ann Rheum Dis 2007;66:1485–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17472993. [Accessed 29 July 2016].
[22] Klooster PM ten, Vonkeman HE, Voshaar MAHO, et al. Experiences of gout-related disability from the patients' perspective: a mixed methods study. Clin Rheumatol 2014;33:1145–54. Available at: http://link.springer.com/10.1007/s10067-013-2400-6. [Accessed 17 April 2019].
[23] Psychologist SM-A. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. ERIC; 1995. Available at: https://eric.ed.gov/?id=EJ517194. [Accessed 8 January 2019].
[24] Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 2013;22:1889–905. Available at: http://link.springer.com/10.1007/s11136-012-0344-y. [Accessed 28 December 2016].
[25] 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. Available at: http://doi.wiley.com/10.1111/bmsp.12078, . [Accessed 9 January 2019].
[26] van der vignette A. A-R package, org U. Available at: https://pure.uvt.nl/portal/files/1489049/MTO_Van_der_Ark_rapport_GettingStartedWithMokken_2011.pdf. [Accessed 9 January 2019].
[27] Schmitt TA. Current methodological considerations in exploratory and confirmatory factor Analysis. J Psychoeduc Assess 2011;29:304–21. Available at: http://jpa.sagepub.com. [Accessed 16 January 2019].
[28] Henson RK, Roberts JK. Use of exploratory factor Analysis in published research. Educ Psychol Meas 2006;66:393–416. Available at: http://journals.sagepub.com/doi/10.1177/0013164405282485. [Accessed 16 January 2019].
[29] Reise SP. The rediscovery of bifactor measurement models. Multivar Behav Res 2012;47:667–96. Available at: http://www.tandfonline.com/doi/abs/10.1080/00273171.2012.715555. [Accessed 16 January 2019].
[30] DeMars CE. A tutorial on interpreting bifactor model scores. Int J Test 2013;13:354–78. Available at: http://www.tandfonline.com/doi/abs/10.1080/15305058.2013.799067. [Accessed 16 January 2019].
[31] Terwee CB, Bot SDM, Boer MR de, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42. Available at: https://www.researchgate.net/profile/Caroline_Terwee/publication/6637555_Quality_criteria_were_proposed_for_measurement_properties_of_health_status_questionnaires/links/0c960515006c8237c6000000.pdf.
[32] Huysamen GK. Coefficient alpha: unnecessarily ambiguous; unduly ubiquitous. SA J Ind Psychol 2006;32. Available at: http://sajip.co.za/index.php/sajip/article/view/242. [Accessed 25 January 2019].
[33] Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 2009;74:107–20. Available at: http://link.springer.com/10.1007/s11336-008-9101-0. [Accessed 9 November 2017].
[34] Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: comments on sijtsma. Available at: https://link.springer.com/content/pdf/10.1007/s11336-008-9102-z.pdf. [Accessed 24 January 2019].
[35] Nunnally JC. Psychometric theory. second ed. New York: McGraw-Hill; 1978.
[36] Evers A, Lucassen W, Meijer R. COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie). *dare.uva.nl*. Available at: http://dare.uva.nl/document/2/79346. [Accessed 24 January 2019].
[37] Hambleton R, Swaminathan H, Rogers H. Fundamentals of item response theory. 1991.
[38] Thissen D. Reliability and measurement precision. In: Wainer H, editor. Computerized adaptive testing: a primer. second ed. NJ, US: Lawrence Erlbaum Associates Publishers; 2000. p. 159–84.
[39] Rutherford C, Costa D, Mercieca-Bebber R, et al. Mode of administration does not cause bias in patient-reported outcome results: a meta-analysis. Qual Life Res 2016;25:559–74. Available at: http://link.springer.com/10.1007/s11136-015-1110-8. [Accessed 17 January 2019].
[40] Bjorner JB, Rose M, Gandek B, et al. Difference in method of administration did not significantly impact item response: an IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. Qual Life Res 2014;23:217–27. Available at: http://link.springer.com/10.1007/s11136-013-0451-4. [Accessed 17 January 2019].
[41] Labott SM, Johnson TP, Fendrich M, et al. Emotional risks to respondents in survey research. J Empir Res Hum Res Ethics 2013;8:53–66. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24169422. [Accessed 18 January 2019].
[42] Kincaid Jr RF, Rogers R, Chissom B. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975. Available at: http://www.dtic.mil/docs/citations/ADA006655. [Accessed 17 July 2017].
[43] DeWalt DA, Rothrock N, Yount S, et al. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45:S12–21.
[44] Hak T, Methods K, van der V-SR. The Three-Step Test-Interview (TSTI): an observation-based method for pretesting self-completion questionnaires. *ojs.ub.uni-konstanz.de*. 2008. Available at: https://ojs.ub.uni-konstanz.de/srm/article/view/1669. [Accessed 11 March 2018].

[45] Oude Voshaar MA, Klooster PM Ten, Taal E, et al. Dutch translation and cross-cultural adaptation of the PROMIS® physical function item bank and cognitive pre-test in Dutch arthritis patients. Arthritis Res Ther 2012;14:R47. Available at: http://arthritis-research.biomedcentral.com/articles/10.1186/ar3760. [Accessed 11 March 2018].
[46] Boomershine CS. A comprehensive evaluation of standardized assessment tools in the diagnosis of fibromyalgia and in the assessment of fibromyalgia severity. Pain Res Treat 2012;2012:653714. Available at: http://www.hindawi.com/journals/prt/2012/653714/. [Accessed 18 January 2019].
[47] Verhoeven AC, Boers M, van der Linden S. Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis. Ann Rheum Dis 2000;59:966–74. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1753042&tool=pmcentrez&rendertype=abstract. [Accessed 1 June 2015].
[48] Bailly F, Fautrel B, Gossec L. Pain assessment in rheumatology - how can we do better? A literature review. Jt Bone Spine 2016;83:384–8. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1297319X16000026. [Accessed 18 January 2019].
[49] Klooster PM Ten, Drossaers-Bakker KW, Taal E, et al. Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. Pain 2006;121:151–7. Available at: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-32944465827%7B&%7DpartnerID=40%7B&%7Drel=R8.0.0.
[50] Klooster PM Ten, Vonkeman HE, Oude Voshaar MAH, et al. Predictors of satisfactory improvements in pain for patients with early rheumatoid arthritis in a treat-to-target study. Rheumatology (Oxford) 2015;54:1080–6. Available at: https://academic.oup.com/rheumatology/article-lookup/doi/10.1093/rheumatology/keu449. [Accessed 22 January 2019].
[51] Saleh KJ, Davis A. Measures for pain and function assessments for patients with osteoarthritis. J Am Acad Orthop Surg 2016;24:e148–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27755265. [Accessed 18 January 2019].
[52] Haywood KL, Garratt AM, Dawes PT. Patient-assessed health in ankylosing spondylitis: a structured review. Rheumatology (Oxford) 2005;44:577–86. Available at: http://academic.oup.com/rheumatology/article/44/5/577/2899257/Patientassessed-health-in-ankylosing-spondylitis-a. [Accessed 18 January 2019].
[53] Englbrecht M, Tarner IH, van der Heijde DM, et al. Measuring pain and efficacy of pain treatment in inflammatory arthritis: a systematic literature review. J Rheumatol Suppl 2012;90:3–10. Available at: http://www.jrheum.org/cgi/doi/10.3899/jrheum.120335. [Accessed 18 January 2019].
[54] Williams VSL, Smith MY, Fehnel SE. The validity and utility of the BPI interference measures for evaluating the impact of osteoarthritic pain. J Pain Symptom Manag 2006;31:48–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16442482. [Accessed 18 January 2019].
[55] Kapstad H, Rokne B, Stavem K. Psychometric properties of the Brief Pain Inventory among patients with osteoarthritis undergoing total hip replacement surgery. Health Qual Life Outcomes 2010;8:148. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21143926. [Accessed 18 January 2019].
[56] Chen CX, Kroenke K, Stump T, et al. Comparative responsiveness of the PROMIS pain interference short forms with legacy pain measures: results from three randomized clinical trials. J Pain 2018 Jun;20(6):664–75. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1526590018309313. [Accessed 18 January 2019].
[57] Freynhagen R, Baron R, Gockel U, et al. painDETECT: a new screening questionnaire to identify neuropathic components in patients with back pain. Curr Med Res Opin 2006;22:1911–20. Available at: http://www.tandfonline.com/doi/full/10.1185/030079906X132488. [Accessed 18 January 2019].
[58] Mannerkorpi K, Ekdahl C. Assessment of functional limitation and disability in patients with fibromyalgia. Scand J Rheumatol 1997;26:4–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9057795. [Accessed 21 January 2019].
[59] Sun Y, Stürmer T, Günther KP, et al. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee–a review of the literature. Clin Rheumatol 1997;16:185–98. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9093802. [Accessed 21 January 2019].
[60] Pua Y-H, Cowan SM, Wrigley TV, et al. The lower extremity functional scale could be an alternative to the western Ontario and McMaster Universities osteoarthritis index physical function scale. J Clin Epidemiol 2009;62:1103–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19282145. [Accessed 21 January 2019].
[61] Oude Voshaar MAH, Klooster PM ten, Taal E, et al. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes 2011;9:99. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3221621&tool=pmcentrez&rendertype=abstract. [Accessed 10 August 2015].
[62] Oude Voshaar MAH, Klooster PM ten, Glas CAW, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity–driven 20-item short form in rheumatoid arthritis compared with traditional measures. Rheumatology 2015;54:kev265. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26224306. [Accessed 28 December 2016].
[63] Collins NJ, Misra D, Felson DT, et al. Measures of knee function. 2011. Available at: www.sportsmed.org/tabs/research/ikdc.aspx. [Accessed 21 January 2019].
[64] Hoogeboom TJ, de Bie RA, den Broeder AA, et al. The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. BMC Muscoskelet Disord 2012;13:117. Available at: http://bmcmusculoskeletdisord.biomedcentral.com/articles/10.1186/1471-2474-13-117. [Accessed 21 January 2019].
[65] Hewlett S, Hehir M, Kirwan JR. Measuring fatigue in rheumatoid arthritis: a systematic review of scales in use. Arthritis Rheum 2007;57:429–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17394228. [Accessed 21 January 2019].
[66] Oude Voshaar MAH, Klooster PM Ten, Bode C, et al. Assessment of fatigue in rheumatoid arthritis: a psychometric comparison of single-item, multiitem, and multidimensional measures. J Rheumatol 2015;42:413–20. Available at: http://www.jrheum.org/cgi/doi/10.3899/jrheum.140389. [Accessed 31 December 2016].
[67] Lai J-S, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Qual Life Res 2006;15:1179–90. Available at: http://link.springer.com/10.1007/s11136-006-0060-6. [Accessed 17 April 2019].
[68] H Oude Voshaar MA, Vonkeman HE, Courvoisier D, et al. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Life Res 2007;28:187–97. Available at: https://doi.org/10.1007/s11136-018-2007-0. [Accessed 22 January 2019].

[69] Klooster PM ten, Oude Voshaar MA, Gandek B, et al. Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. Health Qual Life Outcomes 2013;11:199. Available at: http://www.hqlo.com/content/11/1/199.
[70] Oude Voshaar MAH, Vonkeman HE, Courvoisier D, et al. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Life Res 2018 Jan;28(1):187–97.
[71] Schalet BD, Revicki DA, Cook KF, et al. Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS(®) physical function. J Gen Intern Med 2015;30:1517–23. Available at: http://link.springer.com/10.1007/s11606-015-3360-0. [Accessed 13 November 2017].
[72] Cook KF, Schalet BD, Kallen MA, et al. Establishing a common metric for self-reported pain: linking BPI pain interference and SF-36 bodily pain subscale scores to the PROMIS pain interference metric. Qual Life Res 2015;24:2305–18. Available at: http://link.springer.com/10.1007/s11136-015-0987-6. [Accessed 22 January 2019].
[73] Cieza A, Geyh S, Chatterji S, et al. ICF linking rules: an update based on lessons learned. J Rehabil Med 2005;37:212–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16024476. [Accessed 19 June 2015].
[74] Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. Eur J Phys Rehabil Med 2018;54:110–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28534606. [Accessed 22 January 2019].
[75] Prodinger B, Tennant A, Stucki G, et al. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. J Health Serv Res Policy 2016;21:223–8. Available at: http://journals.sagepub.com/doi/10.1177/1355819616636411. [Accessed 22 January 2019].
[76] van der Linden W, Glas C. Elements of adaptive testing. 2010.
[77] Bjorner JB, Chang CH, Thissen D, et al. Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res 2007;16:95–108. Available at: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-34447115830%7B&%7DpartnerID=40%7B&%7Drel=R8.0.0.
[78] Holman R, Weisscher N, Glas CAW, et al. The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. Health Qual Life Outcomes 2005;3:83. Available at: http://hqlo.biomedcentral.com/articles/10.1186/1477-7525-3-83. [Accessed 30 December 2016].
[79] Oude Voshaar MAH, Klooster PM ten, Glas CAW, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity–driven 20-item short form in rheumatoid arthritis compared with traditional measures. Rheumatology 2015:kev265. Available at: https://academic.oup.com/rheumatology/article-lookup/doi/10.1093/rheumatology/kev265. [Accessed 23 January 2019].
[80] Hays RD, Spritzer KL, Fries JF, et al. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. Ann Rheum Dis 2015;74:104–7. Available at: http://ard.bmj.com/lookup/doi/10.1136/annrheumdis-2013-204053. [Accessed 23 January 2019].
[81] Wohlfahrt A, Bingham CO, Marder W, et al. Responsiveness of patient reported outcomes measurement information system (PROMIS) measures in RA patients starting or switching a DMARD. Arthritis Care Res (Hoboken) 2018. https://doi.org/10.1002/acr.23617.
[82] Yost KJ, Waller NG, Lee MK, et al. The PROMIS fatigue item bank has good measurement properties in patients with fibromyalgia and severe fatigue. Qual Life Res 2017;26:1417–26. Available at: http://link.springer.com/10.1007/s11136-017-1501-0. [Accessed 23 January 2019].
[83] Katz P, Pedro S, Michaud K. Performance of the patient-reported outcomes measurement information system 29-item profile in rheumatoid arthritis, osteoarthritis, fibromyalgia, and systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2017;69:1312–21. Available at: http://doi.wiley.com/10.1002/acr.23183, . [Accessed 23 January 2019].
[84] Oude Voshaar MAH, Klooster PM ten, Glas CAW, et al. Relative performance of commonly used physical function questionnaires in rheumatoid arthritis and a patient-reported outcomes measurement information system computerized adaptive test. Arthritis Rheum 2014;66:2900–8. Available at: http://doi.wiley.com/10.1002/art.38759, . [Accessed 23 January 2019].
[85] Bartlett SJ, Orbai A-M, Duncan T, et al. Reliability and validity of selected PROMIS measures in people with rheumatoid arthritis. Zhang C, ed. PLoS One 2015;10:e0138543. Available at: http://dx.plos.org/10.1371/journal.pone.0138543. [Accessed 23 January 2019].
[86] Cella D, Lai J-S, Jensen SE, et al. PROMIS fatigue item bank had clinical validity across diverse chronic conditions. J Clin Epidemiol 2016;73:128–34. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0895435616001499. [Accessed 23 January 2019].
[87] Askew RL, Cook KF, Revicki DA, et al. Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. J Clin Epidemiol 2016;73:103–11. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0895435616001451. [Accessed 23 January 2019].
[88] Schalet BD, Hays RD, Jensen SE, et al. Validity of PROMIS physical function measured in diverse clinical samples. J Clin Epidemiol 2016;73:112–8. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0895435616300129. [Accessed 23 January 2019].
[89] Wahl E, Gross A, Chernitskiy V, et al. Validity and responsiveness of a 10-item patient-reported measure of physical function in a rheumatoid arthritis clinic population. Arthritis Care Res (Hoboken) 2017;69:338–46. Available at: http://doi.wiley.com/10.1002/acr.22956, . [Accessed 23 January 2019].
[90] Ameringer S, Elswick RK, Menzies V, et al. Psychometric evaluation of the patient-reported outcomes measurement information system fatigue-short form across diverse populations. Nurs Res 2016;65:279–89. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006199-201607000-00004. [Accessed 23 January 2019].
[91] Merriwether EN, Rakel BA, Zimmerman MB, et al. Reliability and construct validity of the patient-reported outcomes measurement information system (PROMIS) instruments in women with fibromyalgia. Pain Med 2017;18:1485–95. Available at: https://academic.oup.com/painmedicine/article-lookup/doi/10.1093/pm/pnw187. [Accessed 23 January 2019].

[92] Nikolaus S, Bode C, Taal E, et al. Construct validation of a multidimensional computerized adaptive test for fatigue in rheumatoid arthritis. Courvoisier DS, ed. PLoS One 2015;10:e0145008. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26710104. [Accessed 23 January 2019].
[93] Van der Linden W. Linear models for optimal test design. 2006. Available at: https://books.google.nl/books?hl=nl&lr=&id=uAd1eYJniBQC&oi=fnd&pg=PA1&dq=Linear+Models+for+Optimal+Test+Design&ots=H0eIYOFLx9&sig=ReLSfJc4Z0HA8AC1Omgqa57BIyM. [Accessed 27 May 2016].
[94] Oude Voshaar MAH, Klooster P Ten, Vonkeman HE, et al. Rasch measurement in rheumatoid arthritis: deriving psychometrically optimal measures from the Rasch Everyday Activity Limitation item bank. Rheumatology (Oxford) 2018;57:1761–8. Available at: https://academic.oup.com/rheumatology/article/57/10/1761/5043200. [Accessed 17 January 2019].