4 Patient-Reported Outcome Measures Available for Adult Lumbar Scoliosis

Vadim Goz, Joseph F. Baker, and Darrel S. Brodke

Introduction

Patient-reported outcome measures (PROMs) or patient-reported outcomes (PROs) are a set of tools that quantify health states by patient self-report. Traditionally these tools have focused on quantification of pain and function, as the improvement in these two qualities represents key goals consistent across musculoskeletal care. Over the past two decades, PROs have played an increasingly important role in healthcare and particularly in adult spine surgery. The tools available for assessment of pain, function and mental health have undergone a rapid evolution.

Early outcome tools were developed using classical test theory (CTT); these tools will be referred to as legacy measures throughout this chapter. Legacy measures include general assessments of pain and function, such as the Short Form 36 (SF-36) and the Sickness Impact Profile (SIP), and disease-specific measures, such as the Oswestry Disability Index (ODI), which is specific to lumbar spine pathology, and the Scoliosis Research Society (SRS) questionnaire for assessing several domains in patients with spinal deformity. Outcome tools took a major step forward with the development of Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS is a novel tool that has been demonstrated to outperform many legacy measures in spine patients.

PROs play an integral part in research by facilitating comparison of outcomes between interventions, as well as in the pursuit of value in healthcare, and in helping physicians communicate to patients and discuss expectations and outcomes of treatment. This chapter will cover a broad range of topics with regard to PROMs including available tools, methodology used for developing outcome tools, the evolution of PROs and the current and future roles of PROs in orthopaedics.

Legacy Outcome Measures

Legacy outcome measures are a group of tools that have served as the foundation of PROs. There are two general types: general outcome measures and disease-specific measures. General measures allow comparisons of patients' health across different medical conditions, for example, comparison of spinal surgery to cardiac surgery. General considerations for assessing these measurement tools include their validity, reliability and responsiveness. A summary of key terms used for assessing the usefulness of a PROM is shown in Table 4.1.

V. Goz, MD
Department of Orthopaedic Surgery, University of Utah, Salt Lake City, UT, USA

J.F. Baker, FRCS
Deparment of Spine and Spinal Deformity Surgery, NYU Hospital for Joint Diseases, New York, NY, USA

D.S. Brodke, MD (✉)
Department of Orthopaedics, University of Utah, Salt Lake City, UT, USA
e-mail: Darrel.Brodke@hsc.utah.edu

E.O. Klineberg (ed.), *Adult Lumbar Scoliosis*, DOI 10.1007/978-3-319-47709-1_4

Table 4.1 Key terms and definitions

Concurrent validity	A measure is compared to an already established, validated measure
Criterion validity	A measure is compared with a similar variable
Discriminative validity	Refers to a measure's ability to differentiate between the various stages and severities of a disease process
Domain	A single trait or characteristic such as pain, function, social health and mental health. Can be subdivided into related groups of traits (e.g. types of pain) called subdomains
External responsiveness	Ability to detect change as a result of some external modifier, e.g. a change in mental health impacting on physical domain
Internal consistency	This measures whether questions in a particular domain actually represent that domain and is reported using a statistical measure: Cronbach's α
Internal responsiveness	Ability to detect expected change, e.g. improvement or otherwise after surgical intervention
Psychometrics	The science of using quantitative tools to measure skills, knowledge and traits, as well as the science of developing and evaluating those tools
Reliability	A reliable measure is one free from random error
Reproducibility	Also known as test-retest reliability and reported using the intra-class coefficient (ICC). Score approaching one confers greater reliability
Responsiveness	Ability of a measure to detect change over time, i.e. detect treatment effects or the changes according to natural history of the disease
Validity	To validate a measure, it needs to be compared with a known standard or process – there are three types of validity
Trait	A characteristic or skill such as pain, function or mental health
Computer adaptive testing (CAT)	A technique by which the response to a given item determines the next item to be administered to a test taker. This produces a customized test, based on the trait level of the examinee that minimizes the number of questions required for a test to estimate a testee's ability
Unidimensionality	The ability of a test/question to assess a single trait without influence by confounders

Understanding measurement tools is essential to interpret results and outcomes from clinical studies and treatments. As an example, Fairbank highlighted previously a potential flaw in reporting outcome data when a non-validated version of the Oswestry Disability Index was used that, when tested, actually resulted in a much higher baseline score than the contemporary validated version [1]. A general understanding of these measures is key to assessing their utility and limitations in spine patients.

General Measures

Short Form 36

Short Form 36 (SF-36) is one of the most widely used tools to assess a patient's general condition and has been translated into over 40 languages. The Medical Outcomes Study Short Form (SF) questionnaires include 6, 12 and 36 question versions. The shortened forms were developed for ease of use and rapid completion [2, 3]. They are most useful for determining the general health of an individual and are used across a variety of surgical and non-surgical fields.

The SF-36 takes between 5 and 10 min to answer all of the questions, and it assesses eight different domains: physical function, bodily pain, social functioning, general mental health, vitality, role limitations due to physical health, and role limitations due to emotional problems and general health perceptions [4]. It can be used to assess and report outcomes from a single domain (i.e. bodily pain or physical function), or the answers can be rolled up into two combined scores, Physical Component Score (PCS) and Mental Component Score (MCS). It has been shown to be acceptable to patients with moderate disabilities although changes have been suggested to accommodate patients who are wheelchair bound, for example, after spinal cord injury [5, 6]. A strength of the SF-36 is the existence of normative data to allow comparison to the population mean [7]. More details on the SF-36 including a comprehensive review of the literature pertaining to the analysis of the scoring, details of development and application for use of the scoring tools and licencing are available online (http://www.sf-36.org/). A disadvantage is that the SF-36 is copyrighted and a licencing fee is required for its use in commercial applications, though generally for non-commercial applications, a licence can be obtained without a fee.

The SF-12 was developed in 1996 as an abbreviated form of the full survey. It can be recorded in the same mode as the SF-36 but has the advantage of taking less than 5 min to complete. The SF-12 is not as sensitive in detecting change at the level of the individual but is fine as a population tool. It also generally requires a licence to use.

The SF-6D is a preference score or quality metric that utilizes six dimensions from the SF-36 – the general health perceptions were omitted, and the limitations as a result of physical and emotional problems were combined. Brazier et al. also developed it as a utility measure for cost-effectiveness research (CER) [8]. In total it describes 18,000 different health states, and anyone completing the SF-36 and SF-12 can be classified according to the SF-6D. Importantly the SF-6D allows one to obtain quality of life-adjusted years for cost-utility analysis (CUA) (like the EQ-5D, discussed below). The SF-6D is also copyrighted and a licence is required.

A concern with any specific PROM is its ability to represent and detect change in clinical status according to treatment. Condition-specific PROs have been tested against the SF-36. Haro et al. jointly assessed the utility of the Japanese Orthopaedic Association (JOA) score, Oswestry Disability Index, Visual Analogue Scale (VAS) and SF-36 (version 2) in a cohort of patients undergoing surgery for lumbar spine stenosis and found good correlation between the four assessments over 24 months of follow-up [9]. The authors determined that the combination of measures was complimentary and the specific strengths of the SF-36 were its assessment of both physical and psychological well-being. Grevitt et al., in a UK cohort of patients undergoing lumbar discectomy, found high reliability for each component of the SF-36. Additionally, all components of the SF-36 correlated well with more specific measures, including the Oswestry Disability Index, except for the mental health domain [10].

Similarly, in a study assessing patient-reported measures in both neck and back disease, Guilfoyle et al. found that SF-36 physical function and bodily pain domains correlated well with the Roland-Morris Disability Index [11]. They also revealed that VAS pain scores for leg pain were strongly correlated with bodily pain scores. They reported that the relevant domains of the SF-36 were free of floor or ceiling effects; however, recent data reveals a significant floor effect for the physical function domain of the SF-36 in the spine patient population, limiting its usefulness [12]. Ware et al. reported on the SF-12 noting acceptable validity and reliability [13]. The SF-6D has good reliability and validity with a significant floor effect, suggesting that it overpredicts poor health states [8, 14, 15].

Veteran RAND Health Surveys

The Veteran RAND (VR) Health Surveys were developed with the support of the Department of Veteran Affairs. These consist of 36- and 12-item questionnaires to assess health-related quality of life across eight domains, much like the SF-36 and SF-12, but however do not require a fee to use. Licencing is still required. Further details about the V-RAND surveys and information about usage can be found online at http://www.rand.org.

The VR-6D is a utility measure composed of six. It was developed in part because of a concern about floor effects of the SF-6D and also the difficulty converting SF-12 scores into SF-6D [16]. The six domains include physical functioning, physical and mental role limitations, social functioning, pain, mental health and vitality. Similar to the SF-6D health state, the scale ranges from 0 to 1 with 0 equivalent to death and 1 being optimum health. It has been shown that as a utility measure, the VR-6D is comparable to the SF-6D [16]. The questionnaires can be completed face to face or over the telephone. Interestingly it has been noted that recording of scores over the telephone results in higher scores (better health quality) than when done face to face [17].

EQ-5D

The EuroQol Group created a non-disease-specific general health measure in 1987 [18]. Initially members included predominantly European nationalities (Dutch, Finnish, Norwegian, Swedish and British); however, the assessment tool has since become increasingly used globally with development centres located in New Zealand, Zimbabwe and the USA among others [19]. It is frequently used as an outcome tool in national registries [20–22].

The principal aims of the EuroQol Group were to create a standardized instrument that would complement rather than replace existing tools for describing health-related quality of life independent of the medical condition of the individual [19, 23]. Details in the measure are available at http://www.euroqol.org. Use of the instrument requires registration and payment of a fee determined by the EuroQol group.

The EQ-5D comprises 245 health states. These are divided into five dimensions and were originally further divided into three levels of severity (3L): no problem, moderate problem and severe problem. After detection of ceiling effects in some general population cohorts, the questionnaire was revised in 2005 to include five levels (5L): no problems, slight problems, moderate problems, severe problems and extreme problems [24]. The dimensions considered include mobility, self-care, usual activities, pain and anxiety/depression.

The EQ-5D can be completed without face-to-face interaction, making completion at home via postal delivery an option. Data gleaned from the EQ-5D can be delivered in three different fashions: it may be reported as a descriptive profile detailing impairment in each dimension, as a population-based score and as a self-rated perceived health status (based on the visual analogue scale component of the questionnaire) [25]. There is a large reference range available for data comparison from the normal population as well as for different diseases making it a useful tool for comparative analyses [26].

The EQ-5D has been tested for its validity in measuring change in health state after lumbar spine surgery for degenerative conditions. Solberg et al. compared it to the ODI in a cohort of over 300 patients undergoing such surgery with 12 months follow-up [27]. They determined cross-sectional construct validity of the EQ-5D in assessing pain, employment, function and health state when compared to the ODI. Only small differences in responsiveness were noted. In a study of patients with adolescent idiopathic scoliosis, the Scoliosis Research Society-22 score was compared with the EQ-5D for repeatability, reliability, consistency and concurrent validity [28]. The authors concluded that the disease-specific and non-specific questionnaires measure different constructs, as the concurrent validity of the EQ-5D was poor to moderate. One drawback for the EQ-5D is the possibility for a ceiling effect and clustering.

Within the field of spine surgery, the EQ-5D has been commonly used in cost-utility analyses (CUAs) [29]. CUA uses ‘health-state utilities’ as an assessment of health outcomes. A utility score provides a preference-based value for a health state ranging from 0 (death) to 1 (perfect health). In CUA, a common approach to representing health-state utilities has been the quality-adjusted life year (QALY). QALYs are defined as the area under the curve of a graph of health-state utility versus time. The EQ-5D has proven to be a useful tool for defining health-state utility scores from which QALYs can then be calculated.

CUA is particularly useful for evaluating outcomes of care where the intended outcome is improvement in the quality of life. The great majority of spine surgery falls under this category. An example of the use of CUA in spine surgery is the work by Tosteson et al. [30] that evaluated the cost-effectiveness of operative versus nonoperative treatment for patients with lumbar disc herniation using data from the Spine Patient Outcomes Research Trial (SPORT). For each cohort QALYs were calculated by using EQ-5D-derived health-state utility scores at 6 weeks, 3, 12 and 14 months. Direct and indirect costs were calculated. The data showed that the cost per QALY gained with surgery compared to nonoperative treatment ranged from $34,355 to $69,403.

While cost-utility analysis is a powerful tool that has been used to evaluate a number of spine procedures, it has its limitations. CUA is not a useful tool for evaluation of procedures that are meant to prevent the deleterious outcomes of disease progression. For example, spine surgery for adolescent idiopathic scoliosis would likely not show significant improvement in quality of life after surgery comparing to before, since the primary goal of the procedure is to prevent future complications of untreated scoliosis. The same applies to resection of asymptomatic tumours that will not lead to immediate improvement in quality of life but will lead to improved overall survival.

Sickness Impact Profile

The Sickness Impact Profile was developed by Gilson et al. in 1975 and subsequently revised by Bergner et al. in 1981 [31, 32]. The assessment is more time consuming or burdensome, requiring 20–30 min to complete. It assesses patient performance over 14 different domains of function encountered on a daily basis and is available in a number of different languages [8]. The patient completes the SIP by selecting statement that best applies to them on the day of completing the questionnaire. Such statements include ‘I sit much of the day’. An overall score is calculated with a higher score indicating a greater level of dysfunction. It can thus be reported as a total score is by using a single domain.

It has been well tested for validity and reliability [7, 33]. Deyo et al. tested the SIP for validity and reliability in a back pain population and found to have substantial test-retest reliability with change in the appropriate direction according to clinical status [34]. It may be useful in populations that are seriously ill in which other measures may be limited by floor effects [33].

At present it is a less frequently used general outcome measure having been supplanted by the aforementioned measures. Frequently cited reasons for its lack of use are its length and the time required for completion. This has prompted efforts to create an abbreviated version that may be more user-friendly [35]. Internal consistency of the abbreviated form (SIP-68) has shown to be excellent; however, there is an additional concern for a large ceiling effect of the SIP in healthy populations [36].

McGill Pain Questionnaire

Melzack and Torgerson developed the McGill Pain Questionnaire in 1971 at McGill University [37]. This is a patient-completed questionnaire that is used to describe the quality and intensity of a patient’s pain. There are three components to the questionnaire. The first section comprises a list of descriptors for the type of pain the patient is experiencing across 20 groups. Only those descriptors that match the patients pain are selected with each term assigned a numeric rating (higher score more severe). The second section asks how the pain changes with time, and the third uncovers relieving factors. The final section asks questions to determine the severity of the pain. The score is provided 0 (not seen in a patient with pain) to a maximum pain score of 78.

A short form (SF-MPQ) was reported by Melzack in 1987 consisting of 15 descriptors of pain rated from 0 to 3 with the higher score indicating greater severity [38]. This abbreviated version also included a Visual Analogue Scale and

Present Pain Intensity (PPI) index from the standard MPQ. A further revision with expansion of the rating scales to a wider format allowing rating from 0 to 10 was reported in 2009 [39]. Acceptable validity and reliability were confirmed in a non-spine cohort.

Visual Analogue and Numeric Pain Rating Scales

Visual Analogue Scales (VAS) or Numeric Pain Rating Scales (NPRS) are used to measure a variety of symptoms, with pain being the most frequent application. Often these are subdivided into back and leg pain separately when dealing with lumbar spine pathology.

The VAS is typically represented by a line, often 100 mm in length with one end representing no pain and the other end most severe possible pain and scored 0–100. No localizing marks other than at each end are allowed, as they may influence the answer. The patient is asked to mark the line between ends (no pain and the severest possible pain) that represents their pain level. The score is reported in centimetres or millimetres along the line from 0 to 10 or 0 to 100. The NPRS on the other hand is typically an 11-point scale from 0 through to 10, similarly representing no pain through the worst possible pain and scored as whole numbers from 0 to 10.

Ostelo et al. previously reviewed the literature with the aim of providing guidelines regarding the Mean Clinically Important Difference (MCID) on commonly used measures including both VAS and NPRS [40]. They determined that a change of 15 mm and 2 for the VAS and NPRS, respectively, represented the MCIDs and a change of 30 % from baseline was a useful threshold. Parker et al. determined a broader range of MCID when analysing a cohort of patients undergoing transforaminal interbody fusion with the mean MCID for VAS 2.8 cm or 28 mm and 2.1 cm or 21 mm for the back and leg pain, respectively [41]. A change of two points on the NPRS has also been deemed to signify a clinically important change by Childs et al., who followed patients with low back pain treated with physical therapy for a 4-week period [42].

A common criticism of the VAS and NPRS is that it is not necessarily clear whether pain is being measured on a particular day or whether it is being measured in general. It also seems as sensitive to anxiety as it is to pain itself. The impact of other painful conditions cannot be negated such as neuropathy or arthroses affecting the appendicular skeleton. Depression and somatization can also influence these measures.

Lumbar Spine-Specific Scores

Oswestry Disability Index

The Oswestry Disability Index (ODI) was developed in the 1970s, first reported in 1980, and is one of the most widely used tools in assessment of lumbar spine pathology [43, 44]. Its widespread use more than 30 years on since its development is a tribute to its developer. It is now licenced to the Mapi Research Trust.

The latest version of the ODI is 2.1a, the previous versions being modified in response to feedback from medical specialists [45]. The ODI contains ten questions pertaining to daily activities performed over the preceding 4 weeks, each of which had six ordinal responses. All the questions relate to activities that may be affected by lower back pain. Each question is scored from 0 to 5; no interference with said activity to maximal interference. The score is then doubled to provide a percentage score from 0 to 100. Scores from 0 to 10 are considered normal, 11–20 minimally disability, 21–60 significantly and increasingly disabled and 61–80 bedridden, while scores over 80 may be spurious [44]. The MCID has been reported previously as 12.8 in a systematic review of patients with an established surgical pathology [46].

The ODI requires no training to use, is self-administered and can be completed in less than 5 min. It has excellent test-retest reliability and has proven validity. Among subjects considered to be ‘unchanged’, Davidson and Keating reported an ICC of 0.74 [47]. It has been well correlated with the Quebec Back Pain Disability Index [48]. Grevitt et al., as mentioned earlier, have shown excellent correlation of the ODI with the SF-36, particularly the physical component of the general score [10].

Criticisms of the ODI include some difficulty with the phrasing of certain questions particularly when considering North American responders [49]. Some modifications have been made to the original version, but one must be careful to ensure that the modified version used is actually a properly validated version to avoid drawing inaccurate or misleading conclusions about treatment effect. The current correct version is 2.1a, and a side-by-side comparison of this with an unvalidated version can be seen in the *Journal of Neurosurgery: Spine* [45].

In a recent study examining the ODI (v2.0) in comparison to PROMIS, Brodke et al. showed that the ODI physical function domain (PFD) in fact has significantly greater ceiling and floor effects, more so floor [50]. When comparing ODI to both SF-36 and PROMIS, the ODI was also shown to have poorer reliability. When assessing the psychometrics and performance of the ODI (v2.0) in a cohort of over 1600 patients with back pain while reaching the conclusion that the ODI performed relatively well, floor and ceiling effects were again detected limiting interpretation of patients at the ends of the spectrum, and suboptimal unidimensionality was demonstrated (inability to accurately measure a single construct without influence from other variables, e.g. depression or anxiety) [12]. Further discussion on the use of PROMIS and conversion from the ODI to PROMIS is discussed below.

Roland-Morris Disability Questionnaire

In 1983, Roland and Morris, both from a general practice background, published this measure of low back pain to assess disability encompassing a wide range of functional domains [51]. It was tried and tested initially in a general practice cohort with testing on almost 200 subjects at weeks 0, 1 and 4. The original version was developed from the Sickness Impact Profile with modification of the questions to include the phrase 'because of my back' [52]. It contained 24 items, but was later revised to include only 18 [53].

Little training in its used is required and is considered easy to complete taking approximately 5 min [49]. It is widely available, and the original 24-statement version can be obtained free from http://www.srisd.com/Roland-Morris.pdf. Translations are available in several languages. Unlike others there is no determining degree or severity of disability in each of the activities – the patient either has or has no difficulty on the given day. The number of items checked off over time can track improvement. The MCID has been determined to be only 2–3 points or a 30 % reduction in baseline score [54].

It has shown excellent internal consistency. Over 200 patients completed the questionnaire twice within 2–4 days with an ICC of 0.91 [48]. However, Davidson and Keating reported an ICC of 0.53 in almost 50 patients, unchanged in symptomatology, retested after 4 weeks [47]. It has been able to distinguish patients who are working from those who are not and those who require medication for their back condition [48].

While its ease of use and widespread use are positives, its dichotomous response categories are seen as a weakness compared to other measures that offer either multiple responses or a scale to determine degree of severity. Another potential drawback is the lack of psychosocial or psychological disability analysis, and hence there is less correlation with other measures that include these domains.

North American Spine Society (NASS) Lumbar Spine Outcome Assessment Instrument

NASS created a taskforce in 1991 for the purpose of developing an outcome measure for the impact of lumbar spine pathology. Daltroy was the lead author in the creation of this instrument, and they reported their development of this tool in 1996 [55].

It contains 34 items and these are broken down into summative scales. In addition there are a series of single-item questions. The scales include pain and disability, neurogenic symptoms, job difficulty, job exertion, expectations and satisfaction. Each subscale is cored from 1 to 6, best to worst. The mean of all items in each subscale is used as the scale score.

No training is required to use the tool and is a self-administered written questionnaire. It is easily accessed from the American Academy of Orthopaedic Surgeons (AAOS) website (www.aaos.org). Interclass coefficients testing reproducibility of the various subscales were all 0.85 or above [49]. The NASS Pain and Disability Scale has been strongly correlated with Visual Analogue Pain Scale, the SF-36 Pain Scale and the SF-36 Physical Limitation Scale [49].

On the flipside a reading level of eighth grade is required which is higher than the significant portion of the US population [56]. Consideration needs to be given for testing the NASS instrument in non-surgical populations and in longitudinal cohort studies [49].

Lumbar Stiffness Disability Index

This is a more recent addition to the armoury of PROMs for the lumbar spine, created and reported by Hart et al. in 2013 [57]. It was born out of a desire to determine what functional impairment resulted from the loss of movement as a consequence of arthrodesis as opposed to loss from pain and other symptoms.

A ten-item questionnaire was tested for validity, reliability and consistency in a cohort of 32 patients undergoing lumbar spine arthrodesis procedures and followed for a year. The ten items each assess the impact of stiffness on daily activities and result in a score from 0 to 100 with higher scores indicative of greater impairment. The scores were correlated also with the degree of resulting stiffness as determined by the range of movement from T12 to S1 on flexion-extension radiographs of the lumbar spine.

In a later study, it was seen that patients undergoing a single-level arthrodesis actually reported less stiffness according to their LSDI, whereas those who underwent three-, four- or five-level procedures were worse off secondary to the degree of stiffness [58].

Overall, this is a relatively new specific measure but offers assessment of an area that earlier measures have perhaps overlooked. As surgery for adult spinal deformity becomes increasingly utilized, it is likely this measure will have a greater role to play.

Scoliosis Research Society-22

Haher et al. published on the development of the Scoliosis Research Society, the SRS-24, score in 1999 [59]. This was prompted by the lack of patient-reported measures on clinical outcome with a large degree of assessment in the adolescent idiopathic scoliosis population based on radiographic measures.

The initial instrument took approximately 5 min to complete and contained 24 questions. These questions covered seven equally weighted domains: pain, general self-image, post-operative self-image, general function, overall activity level, post-operative function and satisfaction. Reliability was confirmed with a Cronbach's α of over 0.6 for each domain. Test-retest reliability was also confirmed with testing on normal controls.

After concerns regarding test-retest reliability, a modified version was later reported on by Asher et al. having been tested in a cohort of 30 patients who has previously undergone surgery for AIS [60]. The modified version was felt to improve the scope of the instrument but also improve internal consistency. It was comparative to the SF-36 in terms of validity. A single question was later removed due to low internal consistency resulting in the SRS-22, and this version has been well tested for concurrent and discriminatory validity, reliability and responsiveness [61–63].

The latest version SRS-r22 is the result of further minor changes in the function domain [64].

Its utility among adult spinal deformity patients was confirmed by Berven et al. who tested it on 146 patients with scoliosis and 34 without [65]. The SRS-22 had less floor and ceiling effects when compared to the SF-36, and test-retest analysis confirmed a high level of reproducibility – Cronbach's α was over 0.75 for each domain. Bridwell et al. further confirmed its use in the adult population analysing a consecutive series of ASD patients over a 12-month period and comparing the SRS-22 to the SF-12 and ODI [66]. They found the SRS-22 is better equipped to detect change in health status than both the generic measures. Except for pain, each domain retained excellent Cronbach's α scores, and test-retest reliability was excellent. Its responsiveness to change has also been confirmed, particularly in the self-image domain [67]. The reliability and validity of the revised SRS questionnaire have been determined in non-English versions also [68, 69].

Quebec Back Pain Disability Scale

Kopec et al. developed the Quebec Back Pain Disability Scale as a measure of disability secondary to low back pain. As a basis it used the World Health Organizations (WHO) definition of 'disability' as a restriction in performing an important activity. It contains 20 items and utilizes Likert scale responses for each without any breakdown into subscales. It was initially developed across a broad range of subspecialties including family practice and psychiatry as an assessment tool for those with low back pain. A strong positive correlation has been found with the Roland Scale, SF-36 physical function subscale and ODI [49]. It has proven to be a reliable, valid and responsive measure, and its conceptual design linking it with the WHO definition of disability is attractive [49].

No training is required for its use and it takes less than 5 min to complete. No equipment is needed, is considered easy to complete and is available free of charge from the authors.

Test-retest stability was initially thought to be good, with an ICC of 0.89 in subjects who had stable symptoms [47]. Reassessed in a separate study, the ICC dropped to 0.55 in patients who reported no improvement over a 4-week period [70]. Those who were unable to return to employment fared worse than those who were able to return [70]. Kopec et al. also tested the Quebec Back Pain Disability Scale in a cohort of almost 250 patients with back pain over a period of 6 months [48]. Retesting was performed after several days then again after 2–6 months. Test-retest reliability was again high (0.92) and Cronbach's-α was 0.96. Expected changed with time were seen confirming its suitability for detecting change with treatment and the natural evolution of a condition.

Zurich Claudication Questionnaire

This is a self-reported measure that is used most often in clinical trials or studies reporting outcomes for treatment of spinal stenosis. It was first reported as a measure in 1996 to complement existing general health measures [71]. It is also at times referred to as the Swiss Spinal Stenosis Questionnaire. Its validity and test-retest reliability have been confirmed in English and other languages [71–73].

The questionnaire consists of three subscales: symptom severity (seven questions), physical function (five questions) and treatment satisfaction (six questions). Symptom severity scale scores range from 1 to 5, while the remainder range 1–4 with higher scores indicating greater disability or loss of function. All questions relate to the patients perception over the preceding month. The maximum possible score is 79, and the result is typically reported as a percentage of maximum score.

The symptom severity subscale can be broken down into two further sections: a pain domain (questions 1–4) and a neuroischemic domain (5–7). While normally reported in its entirety, the physical function subscale is occasionally reported in isolation. This section asks specifically about walking and activities involving walking and is considered an excellent tool to measure the outcome following treatments for spinal stenosis.

Recently the questionnaire has been used in a number of studies reporting the outcome for interventions for spinal stenosis [74].

Other Specific Scales

A number of other scoring systems exist both non-specific and specific to the spine. A full review is beyond the scope of this chapter. Other systems one may come across include the Waddell Disability index. This was a concise nine-item scoring system used to determine physical disability as a result of back pain [75]. The Million Visual Analogue Scale (VAS) was also developed and reported on in the early 1980s and contained 15 questions each with their own visual analogue scale [76]. The Low Back Outcome Score was designed for patients with back pain and sued weighted questions about a patient's activities (employment, domestic activity, sporting activity, sex life, daily activity, rest), current pain and use of medical services and medication [77].

Classical Test Theory, Item Response Theory and Computer Adaptive Testing: The Evolution of PRO Tools

Legacy measures in orthopaedics were developed using classical test theory (CTT). CTT was originally described in the early twentieth century by Spearman [78]. It involves two key parameters: validity and reliability. The fundamental principle of CTT is that a person's observed score is equal to the true score plus measurement error [79]. In this case both the observed score and the true score are functions of the total score for a given test. A test is then validated in a given population, and the reliability of the test score is specific to the population in which it was validated.

The major limitation is that CTT presumes that a single standard error applies to the entire spectrum of ability covered by the test. In practice the reliability is variable depending on the level of trait being measured. For example, when measuring function, a given test typically is more reliable to differentiate between mid-range function levels and is less reliable at the very high or very low ends of function. In practice, for a test designed using CTT to thoroughly cover the entire spectrum of a trait, it would have to be prohibitively lengthy. The other issue is that a given test is validated as a whole and cannot be modified without revalidation.

Item response theory (IRT) addresses many of the shortcomings of CTT. IRT was developed in the 1920s based on the works of Thurstone and Lord [80]. IRT employs a statistical approach that describes the probability of an individual to answer a single item correctly as being dependent on the difficulty of the item and the trait level of the individual. To simplify this further, if we apply this theory to a math test, it says that the probability of answering a math question correctly depends on how good the testee is at math as well as how difficult the question is. Each item, or question, is individually validated and can be thought of as a single measure or grouped into a set of items to increase precision and coverage. The psychometric properties of a test as a whole are then the sum of the individual properties of each testing item.

The key advantages of IRT modelling over CTT are closely related to IRT's two invariance properties: (1) The properties of a question, such as its ability to estimate a trait, are not dependant on the specific group of patients taking the test. (2) A patient's trait level, such as level of function or pain, is independent of the specific set of questions chosen out of a pool of validated questions [81, 82]. This leads to a number of advantages over CTT when applied to PROs in healthcare.

First, IRT-derived tests can be developed that evaluate domains of health (i.e. physical function or depression) across many disease states, rather than measures specific to one disease. Second, a given test item is an independent tool with predictable properties and measures the same trait with the same difficulty regardless of which other items accompany it. This allows for customized tests with varying items dependent on the level of the trait that needs to be evaluated. In addition, items can be added to the upper end or lower end of the trait scale if needed, to improve coverage.

If a total item bank contains questions that vary from low-function-oriented questions such as “Can you ambulate within the house without an assistive device?” to high-function-oriented questions, such as “Can you run five miles?” this ensures that both the low- and high-functioning individuals are covered and can be accurately assessed by the exam. Patients of widely varying abilities still sit along the same trait scale, just at different locations. Furthermore, a test can be customized to the level of the individual, with higher-functioning individuals getting questions that require a higher trait level and allow for more accurate definition of the test taker’s trait.

The process of selecting appropriate questions to accurately define a test taker’s trait level with minimum number of questions is optimized with computer adaptive testing (CAT). CAT technology utilizes an algorithm to determine which question should follow in a given test based on the response to the prior question(s). For example, if a test taker answers that she can jog 1 mile without difficulty, little additional information will be gained by asking whether she can comfortably ambulate about the household without the use of assistive devices. The test taker’s trait level will be better defined if the next question asks whether she can run 5 miles. This results in significantly less burden on the patient and clinic staff by limiting the total number of questions required to define the test taker’s trait level. Studies show that IRT-derived PRO tools administered using CAT achieve higher levels of accuracy, better coverage of the population and lower burden with many fewer questions than legacy measures developed using CTT [83, 84].

One of the consequences of increased emphasis on value is the increasing importance, and increased support for, comparative effectiveness research (CER). Part of the 2010 Patient Protection and Affordable Care Act emphasizes that clinical care and clinical research must incorporate the patient perspective [85]. PRO tools allow for quantification of both health states by the patient and subsequent comparison of health states before and at different time points after various interventions.

Increased support of CER sets the stage for developing of PROs that measure domain-specific outcomes such as ability to engage in physical activity, depression and sleep quality. These domains have been demonstrated as important to patients and their perception of treatment success [86]. Domain-specific outcomes rely on the theory that health attributes are not disease specific and that each disease state has a unique profile in terms of impact on different health domains. In order for PROs to be successfully integrated into CER, and into clinical practice, these instruments must be carefully calibrated and critically evaluated whether they are able to successfully measure the domains of interest in a timely and efficient manner.

Patient-Reported Outcomes Measurement Information System (PROMIS)

PROMIS began in 2004 as a National Institutes of Health (NIH)-funded initiative to develop a novel outcome tool that has improved precision, reliability and validity as compared to legacy tools developed using CTT and has applicability across a wide range of disease states [87, 88]. This initiative is part of the ‘Roadmap for Clinical Research in the Twenty-First Century’ report presented by the director of the NIH in 2002. The project began as collaboration between six primary research sites, a central core of statisticians and several NIH institutes.

Initial work was focused on developing the PROMIS item library by applying IRT methodology and three key protocols: domain mapping, archival data analysis and qualitative data review. The domain mapping protocol involved domain-specific groups that collaborated to define the domain framework for the PROMIS item bank. The ultimate goal of this framework is to have a number of well organized, when appropriate hierarchical, unidimensional domains that together accurately describe a disease state.

Unidimensionality is the ability of a test or question to assess a single trait without influence by confounders, for example, testing physical function without interference from depression. Each domain group contained experts in the domain-related field as well as statisticians.

The domain framework underwent iterative revisions using literature review, data analysis and consensus opinion to move towards the goal of unidimensional categories that accurately define a disease state. The PROMIS Adult health framework contains four general categories: global health, physical health, mental health and social health [89]. Each of those categories has a number of domains and subdomains under it. For example, the physical health item bank is composed of questions from the following five 'profile' domains: physical function, pain intensity, pain interference, fatigue and sleep disturbance.

The archival data analysis and quality item review (QIR) protocols were used to incorporate questions from existing PRO tools into the PROMIS item banks. Questions from preexisting questionnaires were evaluated and assigned to appropriate domains. Each question underwent extensive psychometric testing via IRT analysis. The QIR protocol carefully examined all questions in each domain and eliminated redundant questions [90]. Large field tests were carried out using IRT methods to calibrate the item bank to the general US population.

The domain-driven approach taken by PROMIS for its item banks is a departure from the disease-specific approach of legacy PRO tools. Domains are unidimensional health attributes, based on the World Health Organization (WHO) domains of health, and the domain-specific approach functions under the assumption that each domain is not unique to a disease. This approach allows for comparison of outcomes between disease states, in patients with various combinations of diseases. The domain-specific approach taken by PROMIS particularly lends itself to comparative effectiveness research [19, 91]. It also may be helpful at the level of individual patient care, for adding an objective measure to the discussion of how the patient is doing with treatment, and may lead to effective shared decision-making.

PROMIS has an ever-expanding number of item banks – currently there are 52 available item banks across the three general domains of mental health, physical health and social health. The physical health domain is perhaps the most helpful domain for spine patient assessment. Under this category, a number of item banks can be useful including physical function, pain interference, pain behaviour and sleep disturbance. Physical function with mobility aids item bank may be particularly useful for older patients that have a lower level of function and use assistive devices for ambulation.

Within the mental health domain, the depression and anxiety item banks offer relevant options. The social health domain offers interesting potential for better understanding spine patients, but has not been looked at yet in this specific population. The 'Ability to Participate in Social Roles and Activities' and 'Satisfaction with Social Roles and Activities' item banks may be particularly applicable to the spine patient population and are worthy of further investigation.

The psychometric properties of the physical function PROMIS item bank have been compared to legacy measures in a number of orthopaedic specialties. PROMIS has been shown to correlate highly with the QuickDASH score in upper extremity but take significantly less time to complete [92, 93]. Tyser et al. found that PROMIS outperformed the QuickDASH in terms of floor and ceiling effects [83]. In the upper extremity, PROMIS was also compared to Constant score, and the Short Musculoskeletal Functional Assessment (SMFA), and was found to correlate highly with all legacy measures while requiring less time to administer [94]. PROMIS outperforms the SMFA in terms of ceiling effect in the trauma population with SMFA ceiling effect of 14 % compared to no measurable ceiling effect for PROMIS [84].

The foot and ankle literature also contains comparisons between PROMIS physical function item bank and legacy measures. PROMIS has better reliability comparing to the Foot and Ankle Ability Measure – Activity of Daily Living (FAAM-ADL) subscale and the Foot Function Index five-point verbal rating scale (FFI-5 pt) and requires less time to administer [95]. PROMIS lower extremity item bank has a better floor and ceiling effect than both the FAAM_ADL and the FFI [96].

The majority of research on PROMIS in the spine literature has also been specific to the physical function (PF) item bank. PROMIS PF CAT has been demonstrated to have impressive ceiling (1.7 %) and floor (0.2 %) effects in a large population of spine patients with diverse range of conditions [97]. Analysis of the Oswestry Disability Index (ODI) in a similar population of spine patients reveals that while it has good reliability (person reliability 0.85, item reliability 1), it has a significant floor effect (29.9 %) and a modest ceiling effect (3.9 %) [98].

Similar findings are seen with analysis of the Neck Disability Index (NDI) with a large floor effect (35.5 %) and significant ceiling effect (4.6 %) [99]. The NDI has other psychometric flaws. It exhibits poor unidimensionality; the unexplained variance of the NDI was 9.4 %. It also has an extremely poor raw score to measure correlation, suggesting that while the scores are ordinal, they are not interval (the distance of between five points at one part of the scale is not the same as the distance between five points at another part of the scale), problematic when discussing MCID or using standard parametric statistics.

Lastly, when contemplating using a new measure, it is important to know if older data can still be used or compared. Score conversion is an important element of the PROMIS system with crosswalk or linking tables developed to convert common general outcome scores to PROMIS measures (http://www.prosettastone.org). Working on correlation of disease-specific measures in the spine, Brodke et al. found that SF-36 and ODI scores can be accurately predicted with the PROMIS PF CAT, allowing for development of linking tables [100].

The Road Ahead: Future Directions of Patient-Reported Outcomes

The next step in the development and utilization of outcomes scores, and PROMIS in particular, is application across the clinical and research settings. There has also been a shift in emphasis from tracking strictly biologic outcomes in clinical trials to tracking more subjective outcomes that patients have identified as important [86]. Patient-reported outcomes are ideally poised to measure the health domains important to patients themselves.

Now, in some settings, patients can fill out PRO measures at home, as well. This creates the possibility of capturing more frequent data points and long-term data points, as well as to collect information prior to the office visit in order to use the data provided by the patient to guide the visit. One of the significant hurdles to meaningful integration of PRO tools into patient care is that while PROs are currently being collected at an increasing rate, data is lacking to support a significant impact on patient care or outcomes [77].

The next step in evolution of PRO tools is to incorporate them into clinical practice. PRO data has the potential to facilitate patient-centred outcome-driven care by providing outcome data to guide informed decision-making both by the patient and the physician. As applications become available that ease the process of viewing aggregated data, physicians can show patients the expected outcomes after various surgical and nonoperative intervention and how a patient is doing compared to their expected

$$\boxed{\text{Value} = \frac{\text{Health outcomes}}{\text{Costs of delivering the outcomes}}}$$

Fig. 4.1 Definition of value

course. This technology will ideally improve patient-physician communication and provide ample evidence on which to base clinical decision-making (Fig. 4.1).

References

1. Fairbank JC. Use and abuse of Oswestry Disability Index. Spine (Phila Pa 1976). 2007;32(25):2787–9.
2. Jenkinson C, Layte R. Development and testing of the UK SF-12 (short form health survey). J Health Serv Res Policy. 1997;2(1):14–8.
3. Jenkinson C, Layte R, Jenkinson D, et al. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? J Public Health Med. 1997;19(2):179–86.
4. Ware Jr JE, Sherbourne CD, The MOS. 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30(6):473–83.
5. Andresen EM, Fouts BS, Romeis JC, Brownson CA. Performance of health-related quality-of-life instruments in a spinal cord injured population. Arch Phys Med Rehabil. 1999;80(8):877–84.
6. Meyers AR, Andresen EM. Enabling our instruments: accommodation, universal design, and access to participation in research. Arch Phys Med Rehabil. 2000;81(12 Suppl 2):S5–9.
7. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmaco Econ. 2000;17(1):13–35.
8. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004;13(9):873–84.
9. Haro H, Maekawa S, Hamada Y. Prospective analysis of clinical evaluation and self-assessment by patients after decompression surgery for degenerative lumbar canal stenosis. Spine J. 2008;8(2):380–4.
10. Grevitt M, Khazim R, Webb J, Mulholland R, Shepperd J. The short form-36 health survey questionnaire in spine surgery. J Bone Joint Surg Br. 1997;79(1):48–52.
11. Guilfoyle MR, Seeley H, Laing RJ. The Short Form 36 health survey in spine disease – validation against condition-specific measures. Br J Neurosurg. 2009;23(4):401–5.
12. Brodke DS, Lawrence BD, Spiker WR, Neese AM, Hung M. PROMIS PF CAT outperforms the ODI and SF-36 physical function domain in 1607 spine patients. Eighth Annual Meeting of the Lumbar Spine Research Society. 9–10 Apr 2015, 2015; Chicago.
13. Ware Jr J, Kosinski M, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34(3):220–33.
14. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271–92.
15. Petrou S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. Health Econ. 2005;14(11):1169–89.
16. Selim AJ, Rogers W, Qian SX, Brazier J, Kazis LE. A preference-based measure of health: the VR-6D derived from the veterans RAND 12-Item Health Survey. Qual Life Res. 2011;20(8):1337–47.
17. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health (Oxf). 2005;27(3):281–91.
18. EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199–208.
19. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001;33(5):337–43.
20. Jansson KA, Nemeth G, Granath F, Jonsson B, Blomqvist P. Health-related quality of life in patients before and after surgery for a herniated lumbar disc. J Bone Joint Surg Br. 2005;87(7):959–64.
21. Stromqvist B. Evidence-based lumbar spine surgery. The role of national registration. Acta Orthop Scand Suppl. 2002;73(305):34–9.
22. Stromqvist B, Jonsson B, Fritzell P, Hagg O, Larsson BE, Lind B. The Swedish National Register for lumbar spine surgery: Swedish Society for Spinal Surgery. Acta Orthop Scand. 2001;72(2):99–106.
23. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37(1):53–72.
24. McCormick JD, Werner BC, Shimer AL. Patient-reported outcome measures in spine surgery. J Am Acad Orthop Surg. 2013;21(2):99–107.
25. Brooks R, Rabin R, de Charro F. The measurement and valuation of health status using EQ-5D: a European perspective. Evidence from the EuroQol BIOMED Research Programme. Kluwer Academic Publishers; 2003.
26. Burstrom B, Fredlund P. Self rated health: is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes? J Epidemiol Community Health. 2001;55(11):836–40.
27. Solberg TK, Olsen JA, Ingebrigtsen T, Hofoss D, Nygaard OP. Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. Eur Spine J. 2005;14(10):1000–7.
28. Adobor RD, Rimeslatten S, Keller A, Brox JI. Repeatability, reliability, and concurrent validity of the scoliosis research society-22 questionnaire and EuroQol in patients with adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2010;35(2):206–9.

29. Angevine PD, Berven S. Health economic studies: an introduction to cost-benefit, cost-effectiveness, and cost-utility analyses. Spine (Phila Pa 1976). 2014;39(22 Suppl 1):S9–15.
30. Tosteson AN, Skinner JS, Tosteson TD, et al. The cost effectiveness of surgical versus nonoperative treatment for lumbar disc herniation over two years: evidence from the Spine Patient Outcomes Research Trial (SPORT). Spine (Phila Pa 1976). 2008;33(19):2108–15.
31. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. Med Care. 1981;19(8):787–805.
32. Gilson BS, Gilson JS, Bergner M, et al. The sickness impact profile. Development of an outcome measure of health care. Am J Public Health. 1975;65(12):1304–10.
33. Lurie J. A review of generic health status measures in patients with low back pain. Spine (Phila Pa 1976). 2000;25(24):3125–9.
34. Deyo RA, Diehl AK. Measuring physical and psychosocial function in patients with low-back pain. Spine (Phila Pa 1976). 1983;8(6):635–42.
35. Nanda U, McLendon PM, Andresen EM, Armbrecht E. The SIP68: an abbreviated sickness impact profile for disability outcomes research. Qual Life Res. 2003;12(5):583–95.
36. Post MW, Gerritsen J, Diederikst JP, DeWittet LP. Measuring health status of people who are wheelchair-dependent: validity of the sickness impact profile 68 and the Nottingham Health Profile. Disabil Rehabil. 2001;23(6):245–53.
37. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain. 1975;1(3): 277–99.
38. Melzack R. The short-form McGill Pain Questionnaire. Pain. 1987;30(2):191–7.
39. Dworkin RH, Turk DC, Revicki DA, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). Pain. 2009;144(1–2):35–42.
40. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1976). 2008;33(1):90–4.
41. Parker SL, Adogwa O, Paul AR, et al. Utility of minimum clinically important difference in assessing pain, disability, and health state after transforaminal lumbar interbody fusion for degenerative lumbar spondylolisthesis. J Neurosurg Spine. 2011;14(5):598–604.
42. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. Spine (Phila Pa 1976). 2005;30(11):1331–4.
43. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy. 1980;66(8):271–3.
44. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976). 2000;25(22):2940–52; discussion 2952.
45. Fairbank JC. Why are there different versions of the Oswestry Disability Index? J Neurosurg Spine. 2014;20(1):83–6.
46. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, medical outcomes study questionnaire short form 36, and pain scales. Spine J. 2008;8(6): 968–74.
47. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther. 2002;82(1):8–24.
48. Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec back pain disability scale. Measurement properties. Spine (Phila Pa 1976). 1995;20(3): 341–52.
49. Katz JN. Measures of adult back and neck function. Arthritis Rheum. 2003;49(5S):S43–9.
50. Brodke DS, Annis P, Lawrence BD, Spiker WR, Neese A, Hung, M. Oswestry Disability Index: a psychometric analysis with 1610 patients. 29th Annual Meeting of the North American Spine Society. 2014; San Francisco.
51. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976). 1983;8(2):141–4.
52. Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv. 1976;6(3):393–415.
53. Stratford PW, Binkley JM. Measurement properties of the RM-18. A modified version of the Roland-Morris Disability Scale. Spine (Phila Pa 1976). 1997;22(20):2416–21.
54. Jordan K, Dunn KM, Lewis M, Croft P. A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. J Clin Epidemiol. 2006;59(1):45–52.
55. Daltroy LH, Cats-Baril WL, Katz JN, Fossel AH, Liang MH. The North American spine society lumbar spine outcome assessment Instrument: reliability and validity tests. Spine (Phila Pa 1976). 1996;21(6):741–9.
56. O'Neill SC, Nagle M, Baker JF, Rowan FE, Tierney S, Quinlan JF. An assessment of the readability and quality of elective orthopaedic information on the Internet. Acta Orthop Belg. 2014;80(2): 153–60.

57. Hart RA, Gundle KR, Pro SL, Marshall LM. Lumbar Stiffness Disability Index: pilot testing of consistency, reliability, and validity. Spine J. 2013;13(2): 157–61.
58. Hart RA, Marshall LM, Hiratzka SL, Kane MS, Volpi J, Hiratzka JR. Functional limitations due to stiffness as a collateral impact of instrumented arthrodesis of the lumbar spine. Spine (Phila Pa 1976). 2014;39(24):E1468–74.
59. Haher TR, Gorup JM, Shin TM, et al. Results of the Scoliosis Research Society instrument for evaluation of surgical outcome in adolescent idiopathic scoliosis. A multicenter study of 244 patients. Spine (Phila Pa 1976). 1999;24(14):1435–40.
60. Asher MA, Min Lai S, Burton DC. Further development and validation of the Scoliosis Research Society (SRS) outcomes instrument. Spine (Phila Pa 1976). 2000;25(18):2381–6.
61. Asher M, Min Lai S, Burton D, Manna B. Scoliosis research society-22 patient questionnaire: responsiveness to change associated with surgical treatment. Spine (Phila Pa 1976). 2003;28(1):70–3.
62. Asher M, Min Lai S, Burton D, Manna B. The reliability and concurrent validity of the scoliosis research society-22 patient questionnaire for idiopathic scoliosis. Spine (Phila Pa 1976). 2003;28(1):63–9.
63. Asher M, Min Lai S, Burton D, Manna B. Discrimination validity of the scoliosis research society-22 patient questionnaire: relationship to idiopathic scoliosis curve pattern and curve size. Spine (Phila Pa 1976). 2003;28(1):74–8.
64. Asher MA, Lai SM, Glattes RC, Burton DC, Alanay A, Bago J. Refinement of the SRS-22 health-related quality of life questionnaire function domain. Spine (Phila Pa 1976). 2006;31(5):593–7.
65. Berven S, Deviren V, Demir-Deviren S, Hu SS, Bradford DS. Studies in the modified Scoliosis Research Society Outcomes Instrument in adults: validation, reliability, and discriminatory capacity. Spine (Phila Pa 1976). 2003;28(18):2164–9; discussion 2169.
66. Bridwell KH, Cats-Baril W, Harrast J, et al. The validity of the SRS-22 instrument in an adult spinal deformity population compared with the Oswestry and SF-12: a study of response distribution, concurrent validity, internal consistency, and reliability. Spine (Phila Pa 1976). 2005;30(4):455–61.
67. Bridwell KH, Berven S, Glassman S, et al. Is the SRS-22 instrument responsive to change in adult scoliosis patients having primary spinal deformity surgery? Spine (Phila Pa 1976). 2007;32(20):2220–5.
68. Lonjon G, Ilharreborde B, Odent T, Moreau S, Glorion C, Mazda K. Reliability and validity of the French-Canadian version of the scoliosis research society 22 questionnaire in France. Spine (Phila Pa 1976). 2014;39(1):E26–34.
69. Schlosser TP, Stadhouder A, Schimmel JJ, Lehr AM, van der Heijden GJ, Castelein RM. Reliability and validity of the adapted Dutch version of the revised Scoliosis Research Society 22-item questionnaire. Spine J. 2014;14(8):1663–72.
70. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry low back pain disability questionnaire and the Quebec back pain disability scale. Phys Ther. 2001;81(2):776–88.
71. Stucki G, Daltroy L, Liang MH, Lipson SJ, Fossel AH, Katz JN. Measurement properties of a self-administered outcome measure in lumbar spinal stenosis. Spine (Phila Pa 1976). 1996;21(7):796–803.
72. Hidalgo Ovejero AM, Menendez Garcia M, Bermejo Fraile B, Garcia Mata S, Forcen Alonso T, Mateo SP. Cross-cultural adaptation of the Zurich Claudication Questionnaire. Validation study of the Spanish version. An Sist Sanit Navar. 2015;38(1):41–52.
73. Yi H, Wei X, Zhang W, et al. Reliability and validity of simplified Chinese version of Swiss Spinal Stenosis Questionnaire for patients with degenerative lumbar spinal stenosis. Spine (Phila Pa 1976). 2014;39(10):820–5.
74. Moojen WA, Arts MP, Jacobs WC, et al. IPD without bony decompression versus conventional surgical decompression for lumbar spinal stenosis: 2-year results of a double-blind randomized controlled trial. Eur Spine J. 2015;24:2295–305.
75. Waddell G, Main CJ. Assessment of severity in low-back disorders. Spine (Phila Pa 1976). 1984;9(2):204–8.
76. Million R, Hall W, Nilsen KH, Baker R, Jayson M. Assessment of the progress of the back-pain patient. Spine. 1982;7(3):204–12.
77. Greenough CG, Fraser RD. Assessment of outcome in patients with low-back pain. Spine (Phila Pa 1976). 1992;17(1):36–41.
78. Spearman C. "General Intelligence" objectively determined and measured. Am J Psychol. 1904;15(2):201–92.
79. Petrillo J, Cano SJ, McLeod LD, Coon CD. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. Value Health. 2015;18(1):25–34.
80. Lord FM. Applications of item response theory to practical testing problems. New York, London; Routledge; 1980.
81. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109–17.
82. Hambleton RK, Jones RW. An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. Edu Meas Issues Pract. 1993;12(3):38–47.
83. Tyser AR, Beckmann J, Franklin JD, et al. Evaluation of the PROMIS physical function computer adaptive test in the upper extremity. J Hand Surg. 2014;39(10):2047–51. e2044.

84. Hung M, Stuart AR, Higgins TF, Saltzman CL, Kubiak EN. Computerized adaptive testing using the PROMIS physical function item bank reduces test burden with less ceiling effects compared with the short musculoskeletal function assessment in orthopaedic trauma patients. J Orthop Trauma. 2014;28(8):439–43.
85. Broderick JE, DeWitt EM, Rothrock N, Crane PK, Forrest CB. Advances in patient-reported outcomes: the NIH PROMIS((R)) measures. EGEMS (Wash DC). 2013;1(1):1015.
86. Kirwan JR, Hewlett SE, Heiberg T, et al. Incorporating the patient perspective into outcome assessment in rheumatoid arthritis--progress at OMERACT 7. J Rheumatol. 2005;32(11):2250–6.
87. Magasi S, Ryan G, Revicki D, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res. 2012;21(5):739–46.
88. Fries J, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol. 2005;23(5):S53.
89. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care. 2007;45(5 Suppl 1):S3.
90. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care. 2007;45(5 Suppl 1):S12.
91. Conway PH, Clancy C. Comparative-effectiveness research—implications of the Federal Coordinating Council's report. N Engl J Med. 2009;361(4):328–30.
92. Döring A-C, Nota SP, Hageman MG, Ring DC. Measurement of upper extremity disability using the patient-reported outcomes measurement information system. J Hand Surg. 2014;39(6):1160–5.
93. Overbeek CL, Nota SP, Jayakumar P, Hageman MG, Ring D. The PROMIS physical function correlates with the QuickDASH in patients with upper extremity illness. Clin Orthop Related Res®. 2015;473(1):311–7.
94. Morgan JH, Kallen MA, Okike K, Lee OC, Vrahas MS. PROMIS physical function computer adaptive test compared with other upper extremity outcome measures in the evaluation of proximal humerus fractures in patients older than 60 years. J Orthop Trauma. 2015;29(6):257–63.
95. Hung M, Baumhauer JF, Brodsky JW, et al. Psychometric comparison of the PROMIS physical function CAT with the FAAM and FFI for measuring patient-reported outcomes. Foot Ankle Int. 2014;35(6):592–9.
96. Hung M, Nickisch F, Beals TC, Greene T, Clegg DO, Saltzman CL. New paradigm for patient-reported outcomes assessment in foot & ankle research: computerized adaptive testing. Foot Ankle Int. 2012;33(8):621–6.
97. Hung M, Hon SD, Franklin JD, et al. Psychometric properties of the PROMIS physical function item bank in patients with spinal disorders. Spine. 2014;39(2):158–63.
98. Brodke DS, Annis P, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1610 patients. Spine J. 2014;11(14):S49.
99. Hung M, Cheng C, Hon SD, et al. Challenging the norm: further psychometric investigation of the neck disability index. The Spine Journal. 2015;15(11): 2440–2445.
100. Brodke DS, Lawrence BD, Ryan Spiker W, Neese A, Hung M. Converting ODI or SF-36 Physical Function Domain Scores to a PROMIS PF Score. Spine J. 2014;14(11):S50.