

Chapter 2

A Guide to PROMs Methodology and Selection Criteria

Maha El Gaafary

Introduction

In general, medical management outcomes can be classified into clinical (e.g., cure, survival), personal (e.g., emotional status, self-helplessness, ability to carry out activities of daily living), and economical (e.g., expenses, cost effectiveness). In a clinical scenario, the outcomes can be clinician reported (e.g., progression of the case in response to therapy), physiologic (e.g., tumor size assessed by ultrasound [US] or magnetic resonance imaging [MRI]), caregiver reported (e.g., functional disability), or patient reported (e.g., symptoms or quality of life) [1, 2]. If the patient is monitored for the outcomes by clinician, caregiver, or researcher, then the outcomes become observer reported outcomes (OROs). On the other hand, if the patient is revealing how he/she feels about their medical problem and its impact on their lives, it becomes patient reported outcome (PRO). Proxy reported outcome is different from a PRO or ORO, as it is a measurement based on the report given by someone else on behalf of the patient or as if he or she is the patient.

As the patient is considered “the center” for any healthcare system, “patient-centered care” got to center stage in discussions of the modern healthcare system [3]. Patient-centered care is considered the best approach able to reflect the quality of personal, professional, and organizational relationships. According to the US Food and Drug Administration (FDA), a patient reported outcome is any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else [4, 5]. No wonder PRO has been used as effectiveness end points in clinical trials as well as in standard clinical practice.

M. El Gaafary (✉)

Community and Public Health Department, Ain Shams University, Cairo, Egypt

e-mail: maha_gafary@yahoo.com

Importance of Patient Reported Outcome Measures

Though medical technology has enabled the treating healthcare professionals to measure the patients' physical, physiological, or biochemical parameters, it is not able to calibrate the treatment outcomes or the disease progression/regression from the patients' perspectives. Further, some data can only be obtained from the patient. This includes [2]:

- Various symptoms
- Symptoms not obvious to observers; e.g., fatigue, headache
- Psychological symptoms; e.g., depression, anxiety
- Symptoms in absence of observer; e.g., sleep disturbances
- Frequency of symptoms; e.g., Does the headache occur daily or weekly or monthly?
- Severity of symptoms; e.g., Headache is severe or moderate or mild?
- Nature and severity of disability of the patient; e.g., How severe is the breathlessness?
- The impact of disease or condition on daily life of the patient; e.g., Does rheumatoid arthritis interfere with the activities of daily living of the patient? If yes, how much is the impact?
- Perception or feeling of the patient toward the disease or the treatment given; e.g., Is the patient satisfied with the treatment given?

Such important symptoms can only be reported by the patient. Patient reported outcome measures (PROMs) represent a formal tool able to make the clinician "treat the patient not just the source of bleeding."

However, the PROMs role goes beyond simple assessment of the patients' symptomatology. Various types of outcomes measured by PROMs include social well-being, cognitive functioning, role activities, personal constructs, satisfaction with care [6], health related quality of life (HRQOL) [7], adherence to medical regimens [8], and clinical trial outcomes [6]. Furthermore, PROMs can be helpful in the determination of the patient's eligibility for certain clinical trials; e.g., inclusion criterion for the trial is female patients with hot flushes. It can be used also for confirmation of the measures; e.g., patients with morning stiffness are most likely to be suffering from rheumatoid disease. In other cases, PROMs can help to interpret the patient's symptoms or eliminate other possibilities; e.g., if the patient has breathlessness and the patient is a smoker then chronic obstructive pulmonary disease (COPD) can be expected rather than anemia. In addition, PROMs are useful for the assessment of patients' compliance or reasons for nonadherence to therapy; for example, are the side effects so severe? Also, PROMs have been used as study end points; e.g., efficacy of analgesic drug by determining pain levels [2]. On the other hand, PROMs play an important role to monitor the case progression and determine its impact on the patient's quality of life. In diseases such as cancer, as a result of the cancer progression patients experience multiple symptoms, economical burden, home management problems, and lack of emotional well-being, all of which can

adversely affect quality of life [9]. The role of QOL/PROMs in cancer care can be considered in the following conditions such as [10]:

- Comparison of two standard therapies having similar survival outcomes
- Identification of negative effects of the therapy when survival time is long
- To find out whether a new therapy is preferable to standard therapy
- To determine whether a therapeutic regimen is better than supportive care only, when survival time is short
 - Identification of the needs for the supportive care
 - Determination of negative effects of the adjuvant therapy
- Targeting problems and making communication easier in clinical practice

PROM Instrument Types

Before selecting a PROMs instrument, clinicians should consider the different tools available and how they meet the requirements of the proposed objectives [6]. Review of the literature revealed seven major types of instruments available. They differ in content as well as their intended purpose or application. In view of the growing interest in the PROMs subject, this classification should not be interpreted too rigidly and is not mutually exclusive:

- *Disease-specific*: e.g., PROMs-Arthritis/Spondyloarthritis/Fibromyalgia, Asthma Quality of Life Questionnaire
- *Population-specific*: e.g., Child health assessment questionnaire, Child Health and Illness Profile-Child Edition/CHIP-CE
- *Dimension-specific*: e.g., Beck Depression Inventory
- *Generic*: e.g., Short-Form Health Survey (SF-36)
- *Individualized*: e.g., Patient Generated Index
- *Summary items*: e.g., UK General Household Survey questions about long-standing illness
- *Utility measures*: e.g., EuroQol, EQ-5D

Disease-Specific/Condition-Specific

These instruments have been developed to measure the patient's perceptions of a specific disease or health problem. The Patient Reported Outcome Measures for Rheumatoid Arthritis consists of eight items that produce four dimension scores relating to activity limitations, quality of life, disease activity measures (pain score, patient global assessment, fatigue, duration of morning stiffness), self-reported joint tenderness, and comorbidity as well as self-helplessness assessment [11].

Being disease-specific, this makes these instruments clinically relevant. On the other side, it is not generally possible to administer disease-specific instruments to individuals without the relevant health problem. This means that health status scores cannot be compared with those for the general population, which is a common approach for assessing the impact of a particular disease on health status. Similarly, it is not possible to make comparisons across treatments for different diseases, which limits the application of disease-specific instruments in economic evaluation where different lines of management for the same condition could be compared.

Population-Specific

In the literature, the term “population-specific” may be used to describe both disease-specific/condition-specific instruments and those specific to particular demographic groups or populations, such as children or elderly people or even culturally specific groups (e.g., Asian, White British).

The Childhood Health Assessment Questionnaire (CHAQ) consists of eight subscales: dressing and grooming, arising, eating, walking, hygiene, reach, grip, and activities in addition to visual analog scale (VAS) for pain and global assessment. There are 30 items in the Disability Index, one item each in the Discomfort Index and Health Status Index. Separate questions are included to assess for aids or devices that the child usually uses for any of the aforementioned activities [12]. The population target is children with juvenile arthritis, 1–19 years of age. CHAQ has been validated for use in children with juvenile idiopathic inflammatory and myopathies. The World Health Organization (WHO) International Classification of Functioning, Disability and Health (ICF) components include impairment (pain), activity limitation (ADLs), and participation restriction-overall health status.

The domains and items of population-specific instruments are mostly more relevant to the group in question; e.g., in the case of young children, a school performance domain is included. A specifically tailored format, such as the use of cartoon and clipart illustrations, is used to convey instructions rather than text. This can make these measures more acceptable and comprehensible—enabling individuals who are often not consulted directly to report on their own health and preferences. However, using population-specific measures carries the same disadvantages as disease-specific measures, ruling out comparisons with the general population, and being difficult to be used to compare the efficacy of particular treatments across population groups.

Dimension-Specific

Dimension-specific instruments assess one particular aspect of health status. Those aspects are summarized as [6]:

- *I—Physical function*: mobility, dexterity, range of movement, physical activity; activities of daily living: ability to eat, wash, dress
- *II—Symptoms*: pain; nausea; appetite; energy, vitality, fatigue; sleep and rest
- *III—Global judgments of health*
- *IV—Psychological well-being*: psychological illness, anxiety, depression; coping, positive well-being and adjustment, sense of control, self-esteem
- *V—Social well-being*: family and intimate relations; social contact, integration, and social opportunities; leisure activities; sexual activity and satisfaction
- *VI—Cognitive functioning*: cognition, alertness, concentration, memory, confusion, ability to communicate.
- *VII—Role activities*: employment, household management, financial concerns
- *VIII—Personal constructs*: satisfaction with bodily appearance, stigma and stigmatizing conditions, life satisfaction, spirituality
- *IX—Satisfaction with care*: The most common types are those that measure psychological well-being.

The Beck Depression Inventory contains 21 items that address symptoms of depression [13]. The instrument was originally developed for use with psychiatric patients but it is increasingly used to assess depression in the physically ill. They provide an assessment of a particular dimension of health that is often more detailed than that provided by disease-specific or generic instruments. There is a wide range of data available for comparing and interpreting results. However, measures of psychological well-being in particular were often developed with a primary objective of discrimination in diagnosis and needs assessment. Therefore, the outcome measures appropriateness of such instruments should be tested carefully before use.

Generic

Generic instruments are designed to measure very broad aspects of health and are therefore potentially suitable for a wide range of patient groups and the general population. The Short-Form Health Survey (SF-36) is one of the most widely used generic instruments [14–16]. It is a 36-item instrument that measures health across eight dimensions of physical functioning, social functioning, role limitations due to physical problems, and role limitations due to emotional problems, mental health, vitality, pain, and general health perceptions. The dimension scores form physical and mental component summary scores [16, 17].

The main advantage of generic instruments is that they are suitable for use across a broad range of health problems. They can be used for comparisons between treatments for different patient groups to assess comparative effectiveness. They can also be used with healthy populations to generate normative data that can be used to compare different patient groups. Their broad scope means that they have potential to capture the influence of comorbidity on health, as well as unexpected positive or negative effects of an intervention. This makes them useful for assessing the impact

of new healthcare technologies when the therapeutic effects are uncertain. However, some level of detail has to be sacrificed, which may limit the relevance of generic instruments when applied to a specific patient population. Generic instruments are also potentially less responsive to clinically important changes in health.

Individualized

Individualized instruments allow respondents to select the content of items and/or rate the importance of individual items. The Patient Generated Index asks respondents to list the five most important areas of their lives affected by a disease or health problem and then to rate how badly affected they are in each area, and in the rest of their lives [18, 19]. They then give a number of “points” to the areas in which they would most value an improvement. The individual area ratings are weighted by the “points” given and summed to produce a single index designed to measure the extent to which a patient’s actual situation falls short of their hopes and expectations in those areas of life in which they most value an improvement.

Individualized instruments address the concerns of the individual patient rather than impose an external standard that may be less relevant. Therefore individualized instruments can have high content validity. However, individualized instruments have to be administered by interview in order to produce response rates similar to those for standardized instruments. This has implications on the feasibility of individualized instruments when compared to standardized instruments that can be self-administered.

Summary Items

Summary items ask respondents to summarize diverse aspects of their health status using a single item or a very small number of items. Since 1974 the General Household Survey for England and Wales has used two questions relating to chronic illness and disability: “Do you have any long-standing illness or disability?” and “Does this illness or disability limit your activities in any way?” Transition items are a form of summary item that ask the respondent to assess their current health compared with a specific point in the past, such as their last clinic visit. The SF-36 contains a transition item that asks: “Compared to one year ago, how would you rate your health in general now: excellent, very good, good, fair, and poor?”

Summary items are brief and make the least demands on respondents’ time. Despite their obvious simplicity, there is some short evidence for the measurement properties of summary items including reliability and validity. Summary items that relate to global health also offer a potential means of exploring apparently contradictory trends in different dimensions of health, for example, an improvement in

physical function that coincides with a deterioration in psychological well-being. However, the brevity of summary items limits the specific inferences that can be made about particular aspects of health. Responses to transition items may suffer from recall bias and may be unduly influenced by current health status.

Utility Measures

Utility measures incorporate preferences or values attached to individual health states and express health states as a single index. This type of instrument produces evidence for the overall value of health states to society and can be used in cost-utility analysis. The EuroQol EQ-5D consists of five items relating to mobility, self-care, main activity, pain/discomfort, and anxiety/depression [20, 21]. On the basis of their responses to the five items, patients are classified into a health state with a preference weight attached. Preferences for health states are derived from general population surveys using techniques such as the rating scale, standard gamble, and time trade-off. These techniques are sometimes used to obtain direct health state values from patients.

Being a single index, this facilitates comparisons between treatments for different health problems and is useful for economic evaluation including cost-utility analysis. Utility measures are usually broad in their focus and are therefore subject to the same criticisms as generic instruments. Some respondents have difficulty understanding the nature of the experimental tasks they are required to perform.

Selecting a PROMs Instrument

Several types of instrument are available in the literature. In selecting an instrument, users must consider the different types of tools that are available and how they meet the requirements of the proposed aim. There are several ways to stratify the PROMs instruments:

- *Type*—The simplest and most useful distinctive approach is to classify them into generic, which can be widely applicable, and those specific to particular health problems or populations. These instruments can be used in a number of applications including clinical trials, economic evaluation, and routine patient care.
- *Mode*—Different modes of instrument administration are presented, the main forms being self-administered and interviewer-administered.
- *Properties*—Instrument selection should be based on a number of criteria including certain psychometric properties such as reliability and validity, as well as more general issues such as the appropriateness of an instrument for a specific application.

- *Evidence*—Instrument selection should consider expert recommendations that are based on comprehensive reviews and professional consensus. The PROMs Bibliography can be searched for reviews and recommendations that relate to specific health problems. However, they are not available for all health problems and often need updating.

Selection Criteria

Following the identification of literature pertaining to instruments it is important that users revise the necessary criteria required to select the most suitable instrument(s). There are eight criteria that should be considered in the selection of patient reported outcome measures. These criteria are not uniformly described in the literature; they are also not prioritized in terms of importance, rather they should be considered in relation to the proposed application and objective.

Appropriateness

Appropriateness is the extent to which instrument content is appropriate to the particular application. “Is the instrument content appropriate to the questions that the application seeks to address?” Is it appropriate or not—this ultimately depends on the users’ specific questions and the content of instruments. Instrument selection is often dominated by **psychometric considerations** of reliability and validity, with insufficient attention given to the content of instruments. The names of instruments and constituent scales or dimensions should not be taken at face value [22]. Users should consider the content of individual items within instruments.

PROMs have three broad measurement objectives: discrimination, evaluation, and prediction [23]:

- *Discrimination* is concerned with the measurement of differences between patients when there is no external criterion available to validate the instrument. For example, measures of psychological well-being have been developed to identify individuals suffering from anxiety and depression.
- *Evaluation* is concerned with the measurement of changes over time. For example, PROMs administered before and after treatment are used as outcome measures in clinical trials.
- *Prediction* is concerned with classifying patients when a criterion is available to determine whether the classification is correct. For example, PROMs may be used in diagnosis and screening as a means of identifying individuals for suitable forms of treatment.

The three measurement objectives are not necessarily mutually exclusive. Discrimination and evaluation may be complementary if both are concerned with the measurement of differences that are clinically important, be they cross-sectional

or longitudinal. However, an item that asks about family history of a particular disease or previous environmental exposure may be useful for determining which patients have the disease (prediction) but will be inappropriate for evaluation.

It is also important to consider how broad a measure of health is required. Specific instruments can have a very restricted focus on symptoms and signs of disease, but may also take account of the impact of disease on quality of life. Generic instruments measure provides broader aspects of health and quality of life that are of general importance. Where feasible, it is recommended that both specific and generic instruments be used to measure health outcomes [24, 25].

Acceptability

Acceptability is the extent to which an instrument is accepted by the patients. Indicators of acceptability include “administration time, response rates, and extent of missing data” [6]. There are a number of factors that can influence acceptability, including the mode of administration, questionnaire design (user friendly), and the health status of respondents. Layout, appearance, and legibility have their effect on whether a responder will either complete or refuse filling out the questionnaire. The format of patient-reported instruments can also influence acceptability. For example, the task faced by respondents completing individualized instruments is usually more difficult than that for instruments based on summated rating scales [19].

The instrument must be presented in a language that is familiar to respondents. Guidelines are available that are designed to ensure a high standard of translation [26, 27]. These guidelines recommend the comparison of several independent translations, back translation, and the testing of acceptability of new translations. Issues of acceptability should be considered at the design stage of instrument and questionnaire development. Patients’ views about a new instrument should be obtained at the pretesting phase, prior to formal tests of instrument measurement properties including reliability [28]. Patients can be asked by means of additional questions or semi-structured interview whether they found any questions difficult or distressing.

Feasibility

Feasibility concerns the ease of administration and processing of an instrument. These are important considerations for staff and researchers who collect and process the information produced by patient-reported instruments [9, 29].

Is the instrument easy to administer and process? Instruments that are difficult to administer and process may impede the conduct of research and disrupt clinical care. The complexity and length of an instrument will have implications on the mode of administration. The mode of administration of the instrument may either complicate or facilitate data collection from the patient. Additional resources are required for interviewer administration over self-administration. Staff training needs must be considered before undertaking interviewer administration. Staff may

also have to be available within the clinic to help patients who have difficulty with self-administration. Finally, staff attitudes and acceptance of patient-reported instruments can make a substantial difference to respondent acceptability.

Interpretability

Interpretability concerns the meaningfulness of scores produced by an instrument. To some extent, the lack of familiarity in the use of instruments may be an obstacle to interpretation. Three approaches to interpretation have been proposed:

1. First, changes in instrument scores have been compared to previously documented change scores produced by the same instrument at, for example, major life events such as loss of a job or with modification in the line of management or lifestyle [10].
2. Secondly, attempts have been made to identify the minimal clinically important difference (MCID), which is equal to the smallest change in instrument scores that is perceived as beneficial by patients [30, 31]. External judgments, including summary items such as health transition questions, are used to determine the MCID.
3. Thirdly, normative data from the general population can be used to interpret scores from generic instruments [32, 33].

The standardization of instrument scores is an extension of this form of interpretation that allows score changes to be expressed in terms of the score distribution for the general population and the deviation in this score with particular types of patients or in particular situations [33].

Precision

How close to the actual patient experience is the instrument measure or score? It relates to methods of scaling and scoring items, and the distribution of items over the range of the construct being measured.

The scaling of items within instruments has important implications for precision. The binary/dichotomous or “yes” or “no” is the simplest form of response category, but it does not allow respondents to report different degrees of difficulty or severity.

The majority of instruments use adjectival or Likert type scales such as strongly agree, agree, uncertain, disagree, and strongly disagree. Visual analog scales appear to offer greater precision but there is insufficient evidence to support this and they may be less acceptable to respondents.

There are a number of instruments that incorporate weighting systems, the most widely used being preferences or values derived from the general public for utility measures such as the EuroQol EQ-5D [20] and the Health Utilities Index [34].

Weighting schemes have also been applied to instruments based on summated rating scales, including the Nottingham Health Profile [35] and the Sickness Impact Profile [36]. Such weighting schemes may seem deceptively precise and should be examined for evidence of reliability and validity.

The items and scores of different instruments may vary in how well they capture the full range of the underlying construct being measured. End effects occur when a large proportion of respondents score at the floor or ceiling of the score distribution. If a large proportion of items have end effects then instrument scores will be similarly affected. End effects are evidence that an instrument may be measuring a restricted range of a construct and may limit both discriminatory power and responsiveness [37, 38].

The application of Item Response Theory (IRT) can further help determine the precision of an instrument. IRT assumes that a measurement construct, such as physical disability, can be represented by a hierarchy that ranges from the minimum to maximum level of disability [39]. IRT has shown that a number of instruments have items concentrated around the middle of the hierarchy with relatively fewer items positioned at the ends [39–41].

The scores produced by such instruments are not only a function of the health status of patients but also the precision of measurement.

Reliability

Reliability refers to whether an instrument is internally consistent or reproducible, and it assesses the extent to which an instrument is free from measurement error. As the measurement error of an instrument increases, this would necessitate an increase in the sample size to obtain precise estimates of the effects of an intervention [6].

Internal consistency is measured with a single administration of an instrument and assesses how well items within a scale measure a single underlying dimension. Internal consistency is usually assessed using Cronbach's alpha, which measures the overall correlation between items within a scale [42].

Caution should be exercised in the interpretation of alpha because its size is dependent on the number of items as well as the level of correlation between items [43].

Reproducibility assesses whether an instrument produces the same results on repeated administrations when respondents have not changed. This is assessed by test-retest reliability. There is no exact agreement about the length of time between administrations, but in practice it tends to be between 2 and 14 days [43].

The reliability coefficient is normally calculated by correlating instrument scores for the two administrations. It is recommended that the intra-class correlation coefficient be used in preference to Pearson's correlation coefficient, which fails to take sufficient account of systematic error [6]. Reliability is not a fixed property and must be assessed in relation to the specific population and context [43].

Validity

Validity is the extent to which an instrument measures what is intended. Validity can be assessed qualitatively through an examination of instrument content, and quantitatively through factor analysis and comparisons with related variables. As with reliability, validity should not be seen as a fixed property and must be assessed in relation to the specific population and measurement objectives.

Content and face validity assess whether items adequately address the domain of interest [6]. They are qualitative matters of judging whether an instrument is suitable for its proposed application. Face validity is concerned with whether an instrument appears to be measuring the domain of interest. Content validity is a judgment about whether instrument content adequately covers the domain of interest.

There is increasing evidence that items within instruments tend to be concentrated around the middle of the scale hierarchy, with relatively fewer items at the extremes representing lower and higher levels of health. Instrument content should be examined for relevance to the application and for adequate coverage of the domain of interest.

Further evidence can be obtained from considering how the instrument was developed. This includes the extent of involvement in instrument development of experts with relevant clinical or health status measurement experience [44].

Validity testing should also involve some quantitative assessment. *Criterion* validity is assessed when an instrument correlates with another instrument or measure that is regarded as a more accurate or criterion variable. Within the field of patient-reported health measurement it is rarely the case that a criterion or “gold standard” measure exists that can be used to test the validity of an instrument. There are two exceptions. The first is when an instrument is reduced in length, with the longer version used as the “gold standard” to develop the short version [16]. Scores for short and long versions of the instrument are compared, the objective being a very high level of correlation. Secondly, instruments that have the measurement objective of prediction have a gold standard available either concurrently or in the future. For example, the criterion validity of an instrument designed to predict the presence of a particular disease (screening) can be assessed through a comparison with the results of diagnosis or a prospective outcome like length of hospital stay or mortality.

In the absence of a criterion variable, validity testing takes the form of *construct* validation. PROMs are developed to measure some underlying construct such as physical functioning or pain. On the basis of current understanding, such constructs can be expected to have a set of quantitative relationships with other constructs. For example, patients experiencing more severe pain may be expected to take more analgesics. Construct validity is assessed by comparing the scores produced by an instrument with sets of variables. Expected levels of correlation should be specified at the outset of studies [45].

Many instruments are multidimensional and measure several constructs, including physical functioning, mental health, and social functioning. These constructs should be considered when assessing construct validity as should the expected relationships with sets of variables. Furthermore, the internal structure of such

instruments can be assessed by methods of construct validation. Factor analysis and principal component analysis provide empirical support for the dimensionality or internal construct validity of an instrument [46]. These statistical techniques can pick up separate health domains within an instrument [47].

Responsiveness

Responsiveness is concerned with the measurement of *important changes* in health and is therefore relevant when instruments are to be used in an evaluative context for the measurement of health outcomes. Does the instrument detect changes over time that matter to patients?

Just as with reliability and validity, estimates of responsiveness are related to applications within specific populations and are not an inherent or fixed property of an instrument.

Responsiveness is usually assessed by examining changes in instrument scores for groups of patients whose health is known to have changed. This may follow an intervention of known efficacy or a specific life event that is known to affect the health aspect measured. Alternatively, patients may be asked how their current health compares to some previous point in time by means of a health transition question. There is no single agreed upon method of assessing responsiveness and a number of statistical techniques are used for quantifying responsiveness.

The effect size statistic is equal to the mean change in instrument scores divided by the baseline standard deviation [48]. The standardized response mean is equal to the mean change in scores divided by the standard deviation of the change in scores [49]. The modified standardized response mean, sometimes referred to as the index of responsiveness, is equal to the mean change in scores divided by the standard deviation of change scores in stable subjects [50]. The denominator for the latter can be derived from the test-retest method of reliability testing.

Ideal Properties of PROMs Instrument

Ideal properties of a PROMs instrument can be summarized as follows [3, 51]:

- It should be specific to the *concept* being measured.
- It should be based on *end-point model*.
- It should have *conceptual equivalence* (equivalence in relevance and meaning across languages and cultures).
- It should be based on the *conceptual framework*.
- It should contain *optimum number of items*.
- It should have *easy and specific measurement properties*; i.e., use of the scale that is the easiest for the intended population to understand.
- It should maintain the *confidentiality* of the patient.
- It should be *reproducible*.

Types of Responses to PROM Items

Responses to a certain PROM item may vary between dichotomous and polytomous scale of measurement. Some item responses are expressed as yes/no, present/absent, and true/false. This is referred to as binary or dichotomous response scale of measurement. However, many tests, questionnaires, and inventories in the behavioral sciences include more than 2 response options. It is a set of answer choices that fall into an order, e.g., from highest to lowest. For example, many personality questionnaires include self-relevant statements (e.g., “I enjoy having conversation with friends”), and respondents are given 3 or more response options (e.g., strongly disagree, disagree, neutral, agree, strongly agree). Such items are known as a *polytomous items*, and they require IRT models that are different from those required by binary items.

A scale may be composed of pictures, numbers, or categories [2]. Recording of events is also one of the methods to determine the response that can be included by the patient, e.g., diary maintain. The following types of response scales or options may be used in a PRO instrument [3].

Types of Rating Scales

Likert Scale

The most frequently used rating scale is the Likert scale. Respondents are offered the choice of selecting 1 of 7 pre-defined or even 9 pre-coded responses, with a neutral point being equivocal along the continuum of the scale. Likert scales may evaluate:

- Agreement (strongly agree, agree, undecided, disagree, strongly disagree),
- Frequency (very frequently, frequently, occasionally, rarely, never),
- Importance (very important, important, moderately important, of little importance, unimportant),
- Likelihood (almost always true, usually true, occasionally true, usually not true, almost never true), or
- Other different attitudes (Fig. 2.1).

Fig. 2.1 Likert scale. Example showing agreement response category options

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

Pictorial Scale

Drawing pictures or clipart can be used to express feelings and emotions items in a PROM instrument. They illustrate the rating scale in a more approachable way, especially within a population with a low level of literacy (Fig. 2.2).

Visual Analog Scale

A psychometric tool measuring that can be used to assess a rather subjective outcome such as feeling pain, happiness, or any characteristic or attitude that cannot be measured in a direct way (Fig. 2.3).

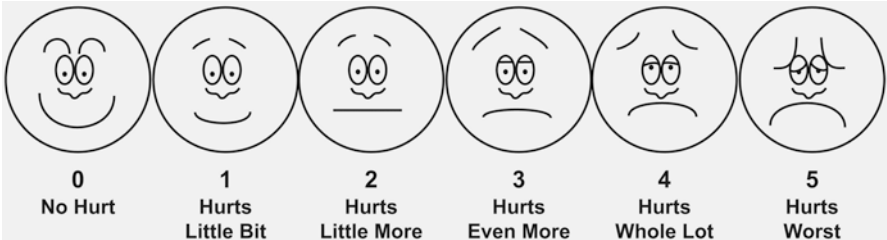


Fig. 2.2 Pictorial scale illustrating different levels of pain as clipart facial expressions

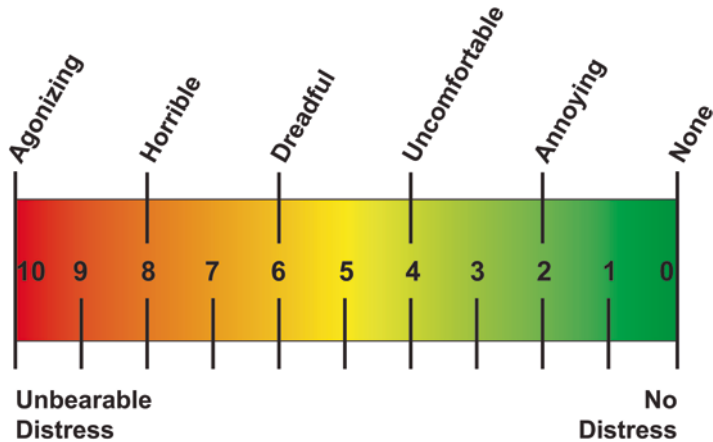


Fig. 2.3 Visual analog scale displaying, on a 10-point scale, the state of distress exerted by a respondent

Rating Scale

A grading continuum is provided for respondents to express the frequency of a particular health event or attack. For example, how many times did you express blue status during the last week? ☐ Once ☐ Twice ☐ Thrice ☐ More than 3

Checklist

Items presented in a PROM using a checklist are usually binary or dichotomous outcome items. Respondents are asked to check in a box the occurrence of certain events or symptoms. There is no response grading or continuum in this case and the item is analyzed as a dichotomous item. Example: Please place a check (✓) in the box in front of the symptoms you expressed more during the last month. ☐ Insomnia ☐ Dyspnea ☐ Body aches ☐ Fatigue

Developing and Validating a PROMS Instrument

Some Important Definitions [52]

PRO Concept—It is the event intended to be measured by the tool. It can be called as the specific measurable goal of the instrument; e.g., symptom or group of symptoms.

PRO Domain—A sub-concept represented by a score of an instrument that measures a larger concept comprised of multiple domains; e.g., depression sometimes referred as scale.

PRO Item—An individual question, statement, or task (and its standardized response options) that should be answered by the patient and it addresses a particular concept; e.g., Are you feeling depressed?

Conceptual Framework—The conceptual framework explicitly defines the concepts measured by the instrument in a diagram that presents a description of the relationships between items, domain (sub-concepts), and concepts measured and the scores produced by a PROMs instrument (Fig. 2.4).

End Point—The measurement that will be statistically compared among treatment groups to assess the effect of treatment or the intervention, and that corresponds again with the intervention's objectives, design, and data analysis.

End-Point Model—A diagram of the hierarchy of relationships among all end points, both PRO and non-PRO, that corresponds to the clinical trial's objectives, design, and data analysis plan (Fig. 2.5).

Conceptual Equivalence—It is the equivalence in relevance and meaning of the concepts being measured in different languages and/or cultures [53].

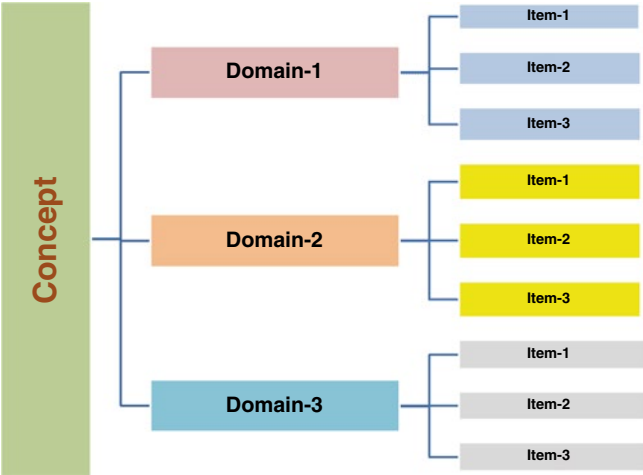


Fig. 2.4 Conceptual framework for a PROM instrument development displaying three domains and corresponding items

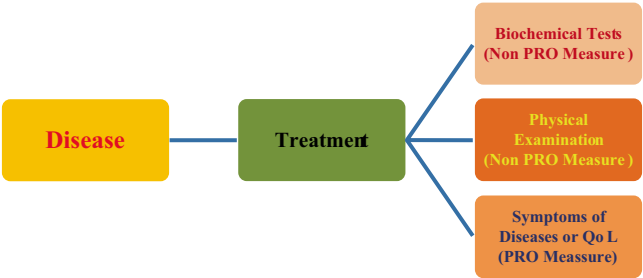


Fig. 2.5 End-point model diagram displaying different outcome measures in response to a specific treatment including PROM

Development of PROMs Instrument

Figure 2.6 illustrates different steps in development of a PROM instrument. The proposed steps are generated from different resources.

Steps in Developing a PROM Instrument

Step I: Conceptual Framework Construction

1. Establishing the need for a new measure by reviewing previous literature.
2. Defining the targeted population in terms of characteristics and accessibility.

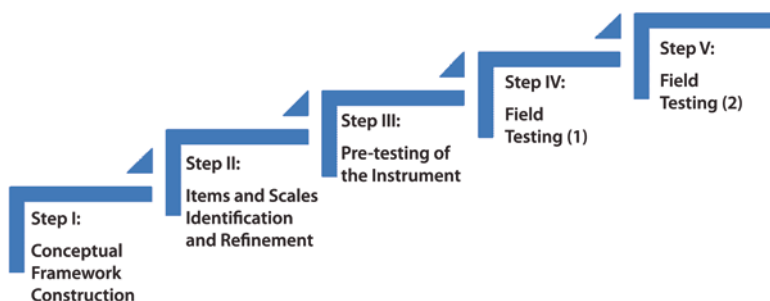


Fig. 2.6 Major steps in developing a PROM instrument as conceptualized from different sources

3. Generating a literature-review domains or scales.
4. Defining the end-point models.
5. Assorting the recall period.
6. Formulating the conceptual framework using clinical expert opinion.
7. Qualitatively interviewing a limited number of people from the concerned population for revision of the preliminary framework.
8. Qualitative analyzing interview transcripts to produce an exhaustive list of items from words taken from people.
9. Generating a preliminary items list.
10. Clinical and psychometric experts review.
11. Production of the conceptual framework of outcomes.

Step II: Items and Scales Identification and Refinement

1. Operationalization and content analysis to classify different items on different domains/scales of measurement.
2. Setting different response options and format.
3. Psychometric analysis with reduction of PROM items using different item selection statistical processes.
4. Clinical and psychometric expert review.
5. Production of the preliminary version of the instrument and its content validity.

Step III: Pretesting of the Instrument

1. Semi-structured cognitive interviews with individuals from the target population—a relatively larger number of respondents are required than step I.
2. Identification of problems with items: ambiguity, confusion, layout and format of the instrument, and the proper mode of administration.
3. Psychometric analysis and revision of the instrument on the base of respondents' recommendations.

4. Rephrasing of items, population response options and mode of administration, translation and cultural adaptation of the instrument, evaluation and documentation of the changes.
5. Clinical and psychometric experts review.
6. Production of the first draft of the version.

Step IV: Field Testing (1)

1. Administration of the instrument to a larger sample of the targeted population—sample size is calculated considering the number of items.
2. Psychometric analysis and modification of the tool according to responses.
3. Testing the effect of mode of administration on differential item functioning (DIF).
4. Rephrasing of items, population response options, and cultural adaptation of the instrument; evaluation and documentation of the changes.
5. Clinical and psychometric experts review.
6. Production of the modified first draft version.

Step V: Field Testing (2)

1. Proper administration of the instrument to a calculated sample of the target population—a control group could be added to test discrimination
2. Final psychometric analysis—traditional and Rasch methods could be used.
3. Minimal modification in rephrasing of items, population response options, and cultural adaptation of the instrument; evaluation and documentation of the changes.
4. Clinical and psychometric experts review.
5. Production of the final version.

Scoring of Items and Domains

For each item, numerical scores should be assigned to each answer category based on the most appropriate scale of measurement for the item (e.g., nominal, ordinal, interval, or ratio scales). Reviewing the distribution of item responses is essential to ensure that response choices represent appropriate intervals. A scoring algorithm creates a single score from multiple items. Equally weighted scores for each item are appropriate when the responses to the items are independent. If two items are dependent, their collected information is less than two independent items and they are over-weighted when they are treated as two equally weighted items. Over-weighting also may be a concern when the number of response options or the values associated with response options vary by item.

Investigators should justify the method chosen to combine items to create a score or to combine domain scores to create a general score using qualitative research or defined statistical techniques.

Total scores combining multiple domains should be supported by evidence that the total score represents a single complex concept. Conceptual framework of a PRO instrument: The instrument's final conceptual framework documents the concept represented by each score [3].

Validation of a PROM Instrument

Validation of different criteria of a PROM instrument requires the use of different psychometric tests and setting off specific criteria would make the instrument ready to be used for a certain objective fulfillment [54].

Acceptability and Data Quality—Completeness of item-level and scale-level data.

- Score distributions: a relatively low number of persons at extreme (i.e., floor/ceiling) ends of the measurement continuum and skewness testing
- Even distribution of endorsement frequencies across response categories (>80 %)
- Percentage of item-level missing data (<10 %)
- Percentage of computable scale scores (>50 % completed items)
- Items in scales rated “not relevant” <35 %

Scaling Assumptions—Legitimacy of summing a set of items (items should measure a common underlying construct).

- Similar items' mean scores and SDs
- Positive residual “ r ” between items (<0.30) to assess model prediction.
- Items have adequate corrected Item to Total Correlation (ITC \geq 0.3).
- High negative residual “ r ” (>0.60) suggests redundancy.
- Items have similar ITCs.
- Items sharing common variance suggest unidimensionality.
- Items do not measure at the same point on the scale.
- Evenly spaced items spanning whole measurement range.

Item Response Categories—Categories are set in a logical hierarchy.

- Ordered set of response thresholds for each scale item.

Targeting—Extent to which the range of the variable measured by the scale matches the range of that variable in the study sample.

- Scale scores spanning entire scale range
- Person-item threshold distribution: person locations should be covered by items.

- Item locations covered by persons when both calibrated on the same metric scale.
- Floor and ceiling (proportion sample at minimum and maximum scale score) effects should be low (<15 %).
- Skewness statistics should range from -1 to +1.
- Good targeting demonstrated by the mean location of items and persons around zero.

Reliability

Internal Consistency—Extent to which items comprising a scale measure the same construct (e.g., homogeneity of the scale).

- Cronbach's alphas for summary scores (adequate scale internal consistency is ≥ 0.70 . Cronbach's α (alpha) is calculated using the following equation [42]:

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^K \sigma_{Y_i}^2}{\sigma_X^2} \right)$$

where K = the number of items

$\sigma(\text{sigma})_X$ = the variance of the observed total test scores

$\sigma(\text{sigma})_{Y_i}$ = the variance of component i for the current sample of persons.

- High person separation index >0.7 ; quantifies how reliably person measurements are separated by items.
- Item-total r (ITC) between +0.4 and +0.6 indicates items are moderately correlated with scale scores; higher values indicate well-correlated items with scale scores.
- Power-of-tests indicate the power in detecting the extent to which the data do not fit the model.
- Items with ordered thresholds.

Test-Retest Reliability—Stability of a measuring instrument.

- Intra-class r coefficient (ICC) >0.70 between test and retest scores
- Statistical stability across time points (no uniform or non-uniform item DIF [$p = >0.05$ or Bonferroni adjusted value])
- Pearson r : >0.7 indicates reliable scale stability

Validity

It involves accumulating evidence from different forms.

Content Validity—Extent to which the content (items) of a scale is representative of the conceptual construct it is intended to measure.

- Consideration of item sufficiency and the target population.
- Clearly defined construct.
- Qualitative evidence from individuals for whom the measure is targeted, expert opinion and literature review (e.g., theoretical and/or conceptual definitions).
- Validity comes from careful item construction and consideration of what each item is meant to measure, then testing against model expectations.

The content validity index (CVI) is widely used for quantifying content validity for scales. Item-level CVI (I-CVI) is calculated by having experts to rate the relevance of each item to its own subdomain (1=not relevant, 2=somewhat relevant, 3=quite relevant, 4=highly relevant). The I-CVI of each item is defined as the number of experts offering a rating of 3 or 4, divided by the total number of experts.

As an adjustment for chance agreements, the multi-rater kappa statistic (K^*) was adopted and is described as follows:

$$P_c = \left[\frac{n!}{A!(n-A)!} \right] \times 0.5^n$$

where P_c is the probability of chance agreement, n is the number of experts, and A is the number approving with good relevance. K^* was calculated using the I-CVI and the probability of chance agreement as follows:

$$K^* = \frac{I - CVI - P_c}{1 - P_c}$$

Each item on the scale was then rated as “fair,” “good,” or “excellent,” based on the following rating criteria: fair, $K^*=0.40$ – 0.59 ; good, $K^*=0.60$ – 0.74 ; excellent, $K^*>0.74$. Any item that received a “fair” rating was deleted [52].

Construct Validity

1. Within-scale analyses

Extent to which a distinct construct is being measured and that items can be combined to form a scale score.

- Cronbach alpha for scale scores >0.70
- Fit residuals (item-person interaction) within given range ± 2.5
- ITC >0.30
- Homogeneity coefficient (IIC mean and range >0.3).

- Nonsignificant chi-square (item-trait interaction) values.
- Scaling success.
- No under- or over-discriminating ICC.
- Mean fit residual close to 0.0; SD approaching 1.0.
- Person fit residuals within given range ± 2.5 .

Measurement Continuum—extent to which scale items mark out the construct as a continuum on which people can be measured.

- Individual scale items located across a continuum in the same way locations of people are spread across the continuum.
- Items spread evenly over a reasonable measurement range. Items with similar locations may indicate item redundancy.
- Response dependency—response to one item determines response to another.
- Response dependency is indicated by residual “ r ” > 0.3 for pairs of items.

2. Between scale analysis

Criterion Validity—hypotheses based on criterion or “gold standard” measure.

- In the majority of cases, there is no true gold standard test for criterion validation of the PROM instrument.

Convergent Validity—scale correlated with other measures of the same/similar constructs.

- Moderate to high “ r ” predicted for similar scales; criteria used as guides to the magnitude of “ r ,” as opposed to pass/fail benchmarks (high $r > 0.7$; moderate $r = 0.3\text{--}0.7$; low $r < 0.3$).

Discriminant Validity—scale not correlated with measures of different constructs

- Low r (< 0.3) predicted between scale scores and measures of different constructs (e.g., age, gender).

Known Groups Differences—ability of a scale to differentiate known groups.

- Generate hypotheses (based on subgroups known to differ on construct measured) and compare mean scores (e.g., predict a stepwise change across severity of illness)
- Hypothesis testing (e.g., clinical questions are formulated and the empirical testing comes from whether or not data fit the Rasch model)
- Statistically significant differences in mean scores (ANOVA)

Differential Item Functioning (Item Bias)—The extent of any conditional relationships between item response and group membership.

- Persons with similar ability should respond in similar ways to individual items regardless of group membership (e.g., age).
- Uniform Differential Item Functioning (DIF)—uniformity amongst differences between groups.
- Non-uniform DIF—non-uniformity amongst differences between groups; can be considered at 1 % (Bonferroni adjusted) and 5 % CIs.

Item Reduction Process

The item-reduction processes of the preliminary scale are resorted to when some items are found to be not relevant or difficult by the qualitative analysis recommendations. They are based on both classical test theory (CTT) (e.g., discrete trend, factor analysis, correlation coefficient, Cronbach's α (alpha) if item deleted [CAID] values, and corrected item-total correlation [CITC]) and item response theory (IRT) [55]. It involves five steps:

1. *Step 1:* Items with low standard deviation indicates low degree of differentiation and should be removed. SD of <0.96 is recommended as a cutoff point.
2. *Step 2:* Principal component factor analysis with varimax rotation to identify the contribution of items to different scales. Sampling adequacy is tested by Kaiser–Meyer–Olkin measure; it should be >0.5 . Items with low factor loading (<0.4) or with factor loading close to other items should be considered for removal.
3. *Step 3:* Item to scale Pearson correlation <0.6 is described as *not* representing the domain or scale.
4. *Step 4:* Internal consistency is evaluated using corrected item to total correlation (CITC) and Cronbach's alpha if item deleted (CAID). CITC >0.45 indicates high contribution of the item to scale, while increased CAID indicates low contribution of the item to scale.
5. *Step 5:* Item Response Theory (IRT) is used in terms of discrimination (α [alpha]) and difficulty (b). Items with α (alpha) <0.4 should be deleted. Items' difficulties are scored on a standardized metric. A range of -3 to $+3$ is allowed. Values for items outside this range are considered for removal.

Both statistical and clinical relevance of items should be taken in account before item removal decision.

IRT and Rasch Models

Item Response Theory is a psychometric approach emphasizing the fact that an individual's response to a particular test item is influenced by qualities of the *individual* and by qualities of the *item*. IRT provides procedures for obtaining information about individuals, items, and tests. Various forms of IRT exist, representing different degrees of complexity or different applicability to various kinds of tests.

The basic form of IRT states that an individual's response to an item is affected by the individual's trait level and the item's difficulty level. More complex forms of IRT include additional factors (or parameters) affecting an individual's responses to items.

Determinants of an Item Response

Respondent Trait Level

One factor affecting an individual's probability of responding in a particular way to an item is the individual's level on the psychological trait being assessed by the item. An individual who has a high level of mathematical ability will be more likely to respond correctly to a math item than will an individual who has a low level of mathematical ability. Similarly, an individual who has a high level of extraversion will be more likely to endorse or agree with an item that measures extraversion than will an individual who has a low level of extraversion. An employee who has a high level of job satisfaction will be more likely to endorse an item that measures job satisfaction than will an employee with a low level of job satisfaction.

Item Difficulty

An item's level of difficulty is another factor affecting an individual's probability of responding in a particular way. A math item that has a high level of difficulty will be less likely to be answered correctly than a math item that has a low level of difficulty (i.e., an easy item).

Similarly, an extraversion measuring item that has a high level of difficulty will be less likely to be endorsed than an extraversion item that has a low level of difficulty. At first, the notion of "difficulty" might not be intuitive in the case of a personality trait such as extraversion, but consider these two hypothetical items: "I enjoy having conversations with friends" and "I enjoy speaking before large audiences." Assuming that these two items are validly interpreted as measures of extraversion, the first item is, in a sense, easier to undertake than the second item. In another way, it is likely that more people would agree with the statement about having a conversation with friends than with the statement about speaking in front of a large audience.

In the context of job satisfaction, the statement "My job is OK" is likely an easier item to agree with than is the statement "My job is the best thing in my life."

Although they are separate issues in an IRT analysis, trait level and item difficulty are intrinsically connected. In fact, item difficulty is conceived in terms of trait level. Specifically, a difficult item requires a relatively high trait level in order to be answered correctly, but an easy item requires only a low trait level to be answered correctly.

In an IRT analysis, trait levels and item difficulties are usually scored on a standardized metric, so that their means are 0 and the standard deviations are 1. Therefore, an individual who has a trait level of 0 has an average level of that trait, and an individual who has a trait level of 1.5 has a trait level that is 1.5 standard deviations above the mean. Similarly, an item with a difficulty level of 0 is an average item, and an item with a difficulty level of 1.5 is a relatively difficult item.

In IRT, item difficulty is expressed in terms of trait level. Specifically, an item's difficulty is defined as the trait level required for participants to have a 0.50 probability of answering the item correctly. If an item has a difficulty of 0, then an individual with an average trait level (i.e., an individual with a trait level of 0) will have a 50/50 chance of correctly answering the item. For an item with a difficulty of 0, an individual with a high trait level (i.e., a trait level greater than 0) will have a higher chance of answering the item correctly, and an individual with a low trait level (i.e., a trait level less than 0) will have a lower chance of answering the item correctly.

Item Discrimination

Just as the items on a test might differ in terms of their difficulties (some items are more difficult than others), the items on a test might also differ in terms of the degree by which they can differentiate individuals who have high trait levels from individuals who have low trait levels. This item characteristic is called item discrimination, and it is analogous to an item–total correlation from classical test theory (CTT) perspectives [56].

An item's discrimination value indicates the relevance of the item to the trait being measured by the test. An item with a positive discrimination value is at least somewhat consistent with the underlying trait being measured, and a relatively large discrimination value (e.g., 3.5 vs. 0.5) indicates a relatively strong consistency between the item and the underlying trait. In contrast, an item with a discrimination value of 0 is unrelated to the underlying trait supposedly being measured, and an item with a negative discrimination value is inversely related to the underlying trait (i.e., high trait scores make it *less* likely that the item will be answered correctly). Thus, it is generally desirable for items to have a large positive discrimination value.

IRT Measurement Models

From an IRT perspective, we can specify the components affecting the probability that an individual will respond in a particular way to a particular item. A *measurement model* expresses the mathematical links between an outcome (e.g., a respondent's score on a particular item) and the components that affect the outcome (e.g., qualities of the respondent and/or qualities of the item).

A variety of models have been developed from the IRT perspective (Table 2.1), and these models differ from each other in at least two important ways. One is in terms of the item characteristics, or *parameters*, that are included in the models. A second is in terms of the response option format.

The simplest IRT model is often called the *Rasch model* or the *one-parameter logistic model* (1PL). According to the Rasch model, an individual's response to a

Table 2.1 Commonly used item response theory (IRT) models

IRT model	Item response format	Model characteristics
Rash/one parameter logistic model	Dichotomous	Discrimination power equal across all items. Threshold varies across items
Two parameters logistic model	Dichotomous	Discrimination and threshold parameters vary across items
Graded response model	Polytomous	Ordered responses. Discrimination varies across items
Nominal model	Polytomous	No pre-specified item response order. Discrimination varies across items
Partial credit model	Polytomous	Discrimination power constrained to be equal across items
Rating scale model	Polytomous	Discrimination equal across items. Distance between item threshold steps equal across items
Generalized partial credit model	Polytomous	Generalization of the partial credit model that allows discrimination to vary across items

binary item (i.e., right/wrong, true/false, agree/disagree) is determined by the individual’s trait level and the difficulty of the item. One way of expressing the Rasch model is in terms of the probability that an individual with a particular trait level will correctly answer an item that has a particular difficulty. This is often presented as [56]:

$$P\left(X_{is} = 1 \mid \theta_s, \beta_i\right) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}$$

where:

- X_{is} refers to response (X) made by subject s to item i .
- $\theta(\theta)_s$ refers to the trait level of subject s .
- $\beta(\beta)_i$ refers to the difficulty of item i .
- $X_{is} = 1$ refers to a “correct” response or an endorsement of the item.
- e is the base of the natural logarithm (i.e., $e = 2.7182818 \dots$), found on many calculators.

So, $P(X_{is} = 1 \mid \theta[\theta]_s, \beta[\beta]_i)$ refers to the probability (P) that subject s will respond to item i correctly or in a particular way. The vertical bar in this statement indicates that this is a “conditional” probability. The probability that the subject will correctly respond to the item depends on (i.e., is conditional upon) the subject’s trait level ($\theta[\theta]_s$) and the item’s difficulty ($\beta[\beta]_i$). In an IRT analysis, trait levels and item difficulties are usually scaled on a standardized metric, so that their means are 0 and the standard deviations are 1.

A slightly more complex IRT model is called the *two-parameter logistic model* (2PL) because it includes 2 item parameters. The difference between the 2PL and

the Rasch model is the inclusion of the item discrimination parameter. This can be presented as [56]:

$$P(X_{is} = 1 | \theta_s, \beta_i, \alpha_i) = \frac{e^{(\alpha_i(\theta_s - \beta_i))}}{1 + e^{(\alpha_i(\theta_s - \beta_i))}}$$

where $\alpha(\alpha)_i$ refers to the discrimination of item i , with higher values representing more discriminating items. The 2PL model states that the probability of a respondent answering an item correctly is conditional upon the respondent's trait level ($\theta(\theta)_s$), the item's difficulty ($\beta(\beta)_i$), and the item's discrimination ($\alpha(\alpha)_i$).

Just as the 2PL model is an extension of the Rasch model (i.e., the 1PL model), there are other models that are extensions of the 2PL model. The *three-parameter logistic model* (3PL) adds yet another item parameter. The third parameter here is an adjustment for guessing. In sum, the 1PL, 2PL, and 3PL models represent IRT measurement models that differ with respect to the number of item parameters that are included in the models.

A second way in which IRT models differ is in terms of the response option format. So far, the 1PL, 2PL, and 3PL models are designed to be used for binary outcomes as the response option. However, many tests, questionnaires, and inventories in the behavioral sciences include more than two response options. For example, many personality questionnaires include self-relevant statements (e.g., "I enjoy having conversation with friends"), and respondents are given three or more response options (e.g., strongly disagree, disagree, neutral, agree, strongly agree). Such items are known as *polytomous items*, and they require IRT models that are different from those required by binary items. Although these models differ in terms of the response options that they can accommodate, they rely on the same general principles as the models designed for binary items. That is, they reflect the idea that an individual's response to an item is determined by the individual's trait level and by item properties, such as difficulty and discrimination.

IRT Models Assumptions [57]:

1. Unidimensionality
2. Local independence
3. IRT model fits the data

It is important that these assumptions be evaluated. However, IRT models are robust to minor violations and no real data ever meet the assumptions perfectly. Unidimensionality requires that the set of items measure a single continuous latent construct $\theta(\theta)$. Scale dimensionality can be evaluated by factor analysis of item responses. If multi-dimensionality is indicated by factor analysis and supported clinical theory, it may be appropriate to divide the scale into subscales.

Local independence means that if $\theta(\theta)$ is held constant, there should be no association among the item responses. Violation of this assumption may result in biased parameter estimated leading to erroneous decisions when selecting items for

scale construction. Local independence can be evaluated by examining the residual correlation matrices for systematic error among item clusters that may indicate violation of the assumption.

Model fit can be examined at both the item and person level to determine whether the estimated item and person parameters can reproduce the observed item responses. Since IRT is probabilistic in nature, most fit indices measure deviations between predicted and observed response frequencies. Many types of residual analysis can be used to evaluate model fit.

Rasch Model(the 1 PL = One Parameter Logistic Model)

The Rasch model includes two determinants of an item response—the respondent's trait level and the items' difficulty level.

The initial estimates of trait levels can be seen as a two-step process. First, we determine the proportion of items that each respondent answered correctly. For a respondent, the proportion correct is simply the number of items answered correctly, divided by the total number of items that were answered. To obtain estimates of trait levels, we next take the natural log of a ratio of proportion correct to proportion incorrect:

$$\theta_s = \text{LN} \left(\frac{P_s}{1 - P_s} \right)$$

where P_s is the proportion of items answered correctly by Respondent 5 (a specific respondent).

The initial estimates of item difficulties also can be seen as a two-step process. First, we determine the proportion of correct responses for each item. For an item, the proportion of correct responses is the number of respondents who answered the item correctly, divided by the total number of respondents who answered the item. To obtain estimates of item difficulty, we compute the natural log of the ratio of the proportion of incorrect responses to the proportion of correct responses:

$$\beta_i = \text{LN} \left(\frac{1 - P_i}{P_i} \right)$$

where P_i is the proportion of correct responses for item i .

Item and Test Information

As a psychometric approach, IRT provides information about items and about tests. In an IRT analysis, item characteristics are combined in order to reflect characteristics of the test as a whole. In this way, item characteristics such as difficulty and

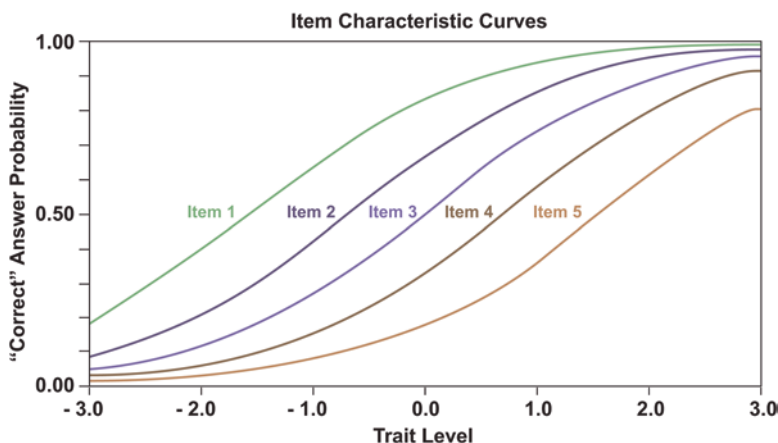


Fig. 2.7 Item characteristic curve for 5 dichotomous items showing different probabilities of a correct answer at different trait levels of respondents [58]

discrimination can be used to evaluate the items and to maximize the overall quality of a test.

Item characteristic curves, such as those presented in Fig. 2.7 [58], reflect the probabilities with which individuals across a range of trait levels are likely to answer each item correctly.

The item characteristic curves in Fig. 2.7 are based on the five items from the hypothetical mathematics test analyzed [58]. For item characteristic curves, the X-axis reflects a wide range of trait levels, and the Y-axis reflects probabilities ranging from 0 to 1.0. Each item has a curve, and we can examine an item's curve to find the likelihood that an individual with a particular trait level will answer the item correctly.

For Item 1, what is the probability that an individual with an average level of mathematical ability will answer the item correctly? We find the point on the Item 1 curve that is directly above the "0" point on the X-axis (recall that the trait level is in z score units, so zero is the average trait level), and we see that this point lies between 0.80 and 0.90 on the Y-axis. Looking at the other curves, we see that an individual with an average level of mathematical ability has about a 0.65 probability of answering Item 2 correctly, a 0.50 chance of answering Item 3 correctly, and a 0.17 probability of answering Item 5 correctly.

By entering an item's difficulty and a particular trait level (say, -3.0) into the model, we obtain the probability with which an individual with that particular trait level will answer that item correctly. We can then enter a different trait level into the model (say, -2.9) and obtain the probability with which an individual with the different trait level will answer the item correctly. After conducting this procedure for

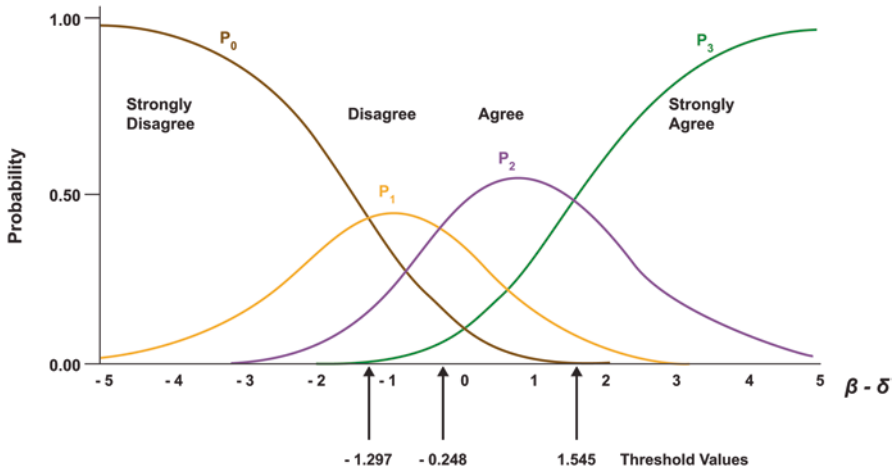


Fig. 2.8 Item characteristic curve for a polytomous item showing the probability of different responses according to different trait levels

many different trait levels, we simply plot the probabilities that we have obtained. The line connecting these probabilities reflects the item's characteristic curve.

Figure 2.8 displays an item characteristic curve for a polytomous item. It shows the highest probability of answering P0 (Strongly Disagree) is associated with the lowest trait level. While P3 (Strongly Agree) corresponds with the highest trait level.

Item information function can identify items that perform well or poorly. Low information for one item may indicate that the item:

1. Measures something different from other items in the scale
2. Is poorly worded and need to be rewritten
3. Too complex for the respondents
4. Placed out of context in the questionnaire

Test Information

From the perspective of CTT, reliability was an important psychometric consideration for a test. For example, we might compute coefficient alpha as an estimate of the test's reliability. An important point to note is that we would compute only 1 reliability estimate for a test, and that estimate would indicate the degree to which observed test scores are correlated with true scores.

The idea that there is a single reliability for a particular test is an important way in which CTT differs from IRT.

From the perspective of IRT, a test does not have a single “reliability.” Instead, a test might have stronger psychometric quality for some people than for others. That is, a test might provide better information at some trait levels than at other trait levels.

How could a test provide information that differs by trait level? Why would a test be able to discriminate between people who have relatively high trait levels but not between people who have relatively low trait levels?

We can use IRT to pinpoint the psychometric quality of a test across a wide range of trait levels. This can be seen as a 2-step process. First, we evaluate the psychometric quality of each item across a range of trait levels. Just as we can compute the probability of a correct answer for an item at a wide range of trait levels (as illustrated in item characteristic curves), we use the probabilities to compute information at the same range of trait levels. For the Rasch model, item information can be computed as follows [56]:

$$I(\theta) = P_i(\theta)(1 - P_i(\theta))$$

where $I(\theta)$ is the item’s information value at a particular trait level (θ), and $P_i(\theta)$ is the probability that a respondent with a particular trait level will answer the item correctly. If we compute information values at many more trait levels, we could display the results in a graph called an *item information curve (IIC)*.

Figure 2.9 illustrates item information characteristics for a 5-item test [58]. It illustrates the spanning of different item information along the trait level of participants in the test.

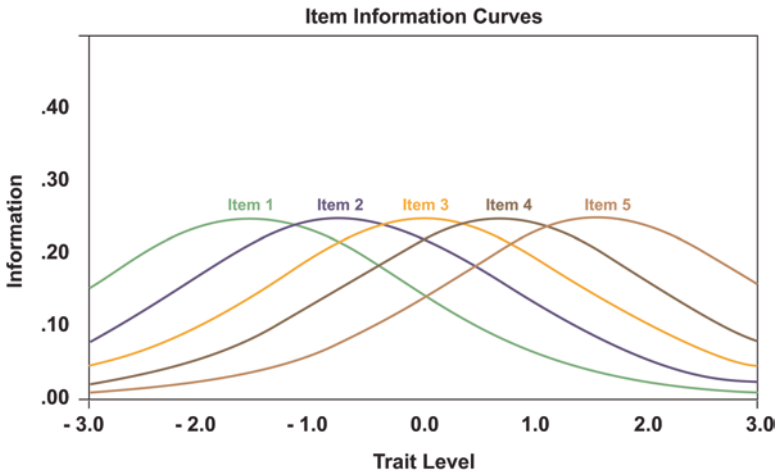


Fig. 2.9 Item Information Curves (IIC) for different items of a test showing different maximum information levels for different items [58]

Higher information values indicate greater psychometric quality. Item 1 has higher psychometric quality at relatively low trait levels than at relatively high trait levels. That is, it is more capable of discriminating among people with low trait levels than among high trait levels (presumably because most people with high trait levels will answer the item correctly).

The height of the curve indicates the amount of information that the item provides. The highest point on a curve represents the trait level at which the item provides the most information. In fact, an item provides the most information at a trait level that corresponds with its difficulty level, estimated earlier. Note that the items differ in the points at which they provide good information. Item 1 provides good information at relatively low trait levels, Item 3 provides good information at average trait levels, and Item 5 provides good information at relatively high trait levels.

Of course, when we actually use a psychological test, we are concerned with the quality of the test as a whole more than the qualities of individual items. Therefore, we can combine item information values to obtain test information values (Fig. 2.10).

Specifically, item information values at a particular trait level can be added together to obtain a test information value at that trait level.

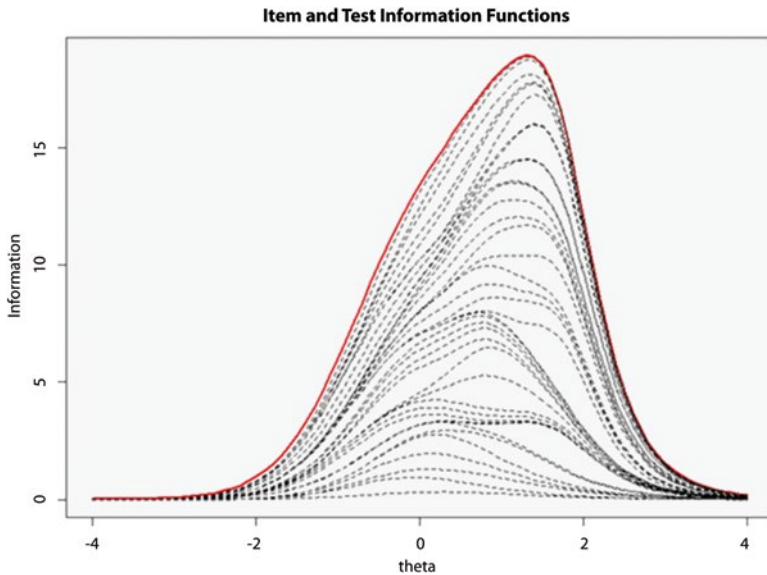


Fig. 2.10 Item and Test Information Function where theta denotes the different ability or trait levels of the respondents

From an IRT perspective, a test's psychometric quality can vary across trait levels. This is an important but perhaps underappreciated difference between the CCT and IRT approaches to test theory.

Differential Item Functioning

From an IRT perspective, analyses can be conducted to evaluate the presence and nature of differential item functioning (DIF). Differential item functioning occurs when an item's properties in one group are different from the item's properties in another group. For example, DIF exists when a particular item has one difficulty level for males and a different difficulty level for females. In another way, the presence of differential item functioning means that a male and a female who have the same trait level have different probabilities of answering the item correctly. The existence of DIF between groups indicates that the groups cannot be meaningfully compared on the item.

For example, Smith and Reise (1998) [59] used IRT to examine the presence and nature of DIF for males and females on the Stress Reaction scale of the Multidimensional Personality Questionnaire. The Stress Reaction scale assesses the tendency to experience negative emotions such as guilt and anxiety, and previous research had shown that males and females often have different means on such scales. Smith and Reise [59] argued that this difference could reflect a true gender difference in such traits or that it could be produced by differential item functioning on such scales. Their analysis indicated that, although females do appear to have higher trait levels of stress reaction, DIF does exist for several items. Furthermore, their analyses revealed interesting psychological meaning for the items that did show DIF. Smith and Reise [59] state that items related to "emotional vulnerability and sensitivity in situations that involve self-evaluation" were easier for females to endorse, but items related to "the general experience of nervous tensions, unexplainable moodiness, irritation, frustration, and being on-edge" were easier for males to endorse. Smith and Reise [59] concluded that inventories designed to measure negative emotionality will show a large gender difference when "female DIF-type items" are overrepresented and that such inventories will show a small gender difference when "male DIF-type items" are overrepresented. Such insights can inform the development and interpretation of important psychological measures.

Person Fit

Another interesting application of IRT is a phenomenon called *person fit* [60]. When we administer a psychological test, we might find an individual whose pattern of responses seems strange compared to typical responses.

Consider 2 items that might be found on a measure of friendliness:

1. I like my friends.
2. I am willing to lend my friends as much money as they might ever want.

Most people would probably agree with the first statement (i.e., it is an “easy” item). In contrast, fewer people might agree with the second statement. Although most of us like our friends and would be willing to help them, not all of us would be willing to lend our friends “as much money as they might ever want.” It would be quite odd to find someone who would be willing to lend any amount of money to her friends if she does not like her friends.

The analysis of person fit is an attempt to identify individuals whose response pattern does not seem to fit any of the expected patterns of responses to a set of items. Although there are several approaches to the analysis of person fit [60], the general idea is that IRT can be used to estimate item characteristics and then to identify individuals whose responses to items do not adhere to those parameters.

The identification of individuals with poor person fit to a set of items has several possible implications. In a personality assessment context, poor person fit might reveal that an individual’s personality is unique in that it produces responses that do not fit the “typically expected” pattern of responses generated from the tested population.

Conclusion

Despite the conceptual and computational challenges and difficulties, the many potential advantages of IRT models should not be ignored. Knowledge of IRT is spreading within the academic disciplines of psychology, education, and public health. More books and tutorials are being written on this subject, and user friendly software is being developed. Research that applies IRT models is appearing more frequently in health outcomes literature. A better understanding of the models and applications of IRT is emerging and this will result in health outcome instruments that are shorter, more reliable, and better at targeting the population of interest.

References

1. Singh DP. Quality of life in cancer patients receiving palliative care. *Indian J Palliat Care*. 2010;16:36–43.
2. Kozma CM, Reeder CE, Schulz RM. Economic, clinical, and humanistic outcomes: a planning model for pharmacoeconomic research. *Clin Ther*. 1993;15:1121–32. Discussion 1120.
3. International Alliance of Patients’ Organizations. What is patient centred health care? A review of definitions and principles. 2nd ed. London: IAPO; 2007. p. 1–34.
4. Chin R, Lee BY. Economics and patient reported outcomes, principles and practice of clinical trial medicine. London: Elsevier; 2008. p. 145–66.
5. U.S. Department of Health and Human Services Food and Drug Administration Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. U.S. FDA, Clinical/Medical. 2009. [<http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf>]. Accessed 17 Aug 2015.

6. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess.* 1998;2:1–74.
7. Chen H, Taichman DB, Doyle RL. Health-related quality of life and patient-reported outcomes in pulmonary arterial hypertension. *Proc Am Thorac Soc.* 2008;5:623–30.
8. Chao J, Nau DP, Aikens JE. Patient-reported perceptions of side effects of antihyperglycemic medication and adherence to medication regimens in persons with diabetes mellitus. *Clin Ther.* 2007;29:177–80.
9. Erickson P, Taeuber RC, Scott J. Operational aspects of quality-of-life assessment: choosing the right instrument: review article. *PharmacoEconomics.* 1995;7:39–48.
10. Testa MA, Simonson DC. Assessment of quality-of-life outcomes. *N Engl J Med.* 1996;334:835–40.
11. El Miedany Y, El Gaafary M, Youssef S, Palmer D. Incorporating patient reported outcome measures in clinical practice: development and validation of a questionnaire for inflammatory arthritis. *Clin Exp Rheumatol.* 2010;28:734–44.
12. Singh G, Athreya B, Fries J, Goldsmith D. Measurement of health status in children with juvenile rheumatoid arthritis. *Arthritis Rheum.* 1994;37(12):1761–9.
13. Beck A, Ward C, Medelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry.* 1961;4:561–71.
14. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): i. Conceptual framework and item selection. *Med Care.* 1992;30:473–83.
15. Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? *Br Med J.* 1993;306:1440–3.
16. Ware JE, Kosinski M, Bayliss MS, McHorney C, Rogers WH, Raczek A. Comparison of methods for scoring and statistical analysis of the SF-36 health profile and summary measures: summary of results from the medical outcomes study. *Med Care.* 1995;33:S264–79.
17. Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey. Construction of scales and preliminary tests of validity and reliability. *Med Care.* 1995;34:220–33.
18. Ruta DA, Garratt AM, Leng M, Russell IT, Macdonald LM. A new approach to the measurement of quality of life: the patient-generated index. *Med Care.* 1994;11:1109–26. 8.
19. Ruta DA, Garratt AM, Russell IT. Patient centred assessment of quality of life for patients with four common conditions. *Qual Health Care.* 1999;8:22–9.
20. Group EQ. EuroQol—a new facility for the measurement of health related quality of life. *Health Policy.* 1990;16:199–208.
21. Brooks R with the EuroQol Group. EuroQol: the current state of play. *Health Policy.* 1996;37:53–72.
22. Ware JE. Standards for validating health measures: definition and content. *J Chronic Dis.* 1987;40:473–80.
23. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis.* 1985;38:27–36.
24. Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality of life measurement: can we keep it simple? *J R Stat Soc.* 1992;155:353–93.
25. Garratt AM, Ruta DA, Abdalla MI, Russell IT. Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. *Qual Life Res.* 1996;5:223–34.
26. Bullinger M. Ensuring international equivalence of quality of life measures. In: Orley J, Kuyken W, editors. *Quality of life assessment: international perspectives.* Berlin: Springer; 1994. p. 33–40.
27. Lepège A, Verdier A. The adaptation of health status measures: methodological aspects of the translation procedure. In: Shumaker S, Berzon R, editors. *The international assessment of health-related quality of life: theory, translation, measurement and analysis.* Oxford: Rapid Communications of Oxford; 1995. p. 93–101.
28. Sprangers MA, Cull A, Bjordal K, Groenvold M, Aaronson NK. The European organization for research and treatment of cancer. Approach to quality of life assessment: guidelines for

- developing questionnaire modules. EORTC study group on quality of life. *Qual Life Res.* 1993;2:287–95.
29. Aaronson NK. Assessing the quality of life of patients in cancer clinical trials: common problems and common sense solutions. *Eur J Cancer.* 1992;28A:1304–7.
 30. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407–15.
 31. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol.* 1994;47:81–7.
 32. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, McGlynn EA, Ware JE. Functional status and well-being of patients with chronic conditions: results from the medical outcomes study. *J Am Med Assoc.* 1989;262:907–13.
 33. Garratt AM, Ruta DA, Abdalla MI, Russell IT. The SF-36 health survey questionnaire: ii responsiveness to changes in health status in four common clinical conditions. *Qual Health Care.* 1994;3:186–92.
 34. Feeny DH, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics.* 1995;7:490–502.
 35. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. *J R Coll Gen Pract.* 1985;35:185–8.
 36. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. *Med Care.* 1981;19:787–805.
 37. Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients: the floor phenomenon. *Med Care.* 1990;28:1142–52.
 38. Gardiner PV, Sykes HR, Hassey GA, Walker DJ. An evaluation of the health assessment questionnaire in long-term follow-up of disability in rheumatoid arthritis. *Br J Rheumatol.* 1993;32:724–8.
 39. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scale and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol.* 1996;49:711–7.
 40. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford health assessment questionnaire. *Br J Rheumatol.* 1996;35:574–8.
 41. Garratt AM in collaboration with UKBEAM. Rasch analysis of the Roland disability questionnaire. *Spine.* 2003;28:79–84.
 42. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951;16:297–334.
 43. Streiner GL, Norman RD. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.
 44. Guyatt GH, Cook DJ. Health status, quality of life and the individual. *J Am Med Assoc.* 1994;272:630–1.
 45. McDowell I, Jenkinson C. Development standards for health measures. *J Health Serv Res Policy.* 1996;1:238–46.
 46. Jolliffe IT, Morgan BJT. Principal component analysis and exploratory factor analysis. *Stat Methods Med Res.* 1992;1:69–95.
 47. Garratt AM, Hutchinson A, Russell IT. The UK version of the Seattle Angina Questionnaire (SAQ-UK): reliability, validity and responsiveness. *J Clin Epidemiol.* 2001;54:907–15.
 48. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care.* 1989;27:MS178–89.
 49. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care.* 1990;28:632–42.
 50. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987;40:171–8.
 51. Acquadro C, Berzon R, Dubois D, Leidy NK, Marquis P, Revicki D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of

- the patient-reported outcomes (PRO) harmonization group meeting at the food and drug administration, February 16, 2001. *Value Health*. 2003;6:522–31.
52. Deshpande PR, Rajan S, Lakshmi Sudeepthi B, Abdul Nazir CP. Patient-reported outcomes: a new era in clinical research. *Perspect Clin Res*. 2011;2(4):137–44.
 53. Pashos CL, Klein EG, Wanke LA, editors. *ISPOR Lexicon™*. Princeton: International Society for Pharmacoeconomics and Outcomes Research; 1998.
 54. Goreck C, Brown JM, Cano S, Lamping DL, Briggs M, Coleman S, Dealey C, McGinnis E, Nelson AE, Stubbs N, Wilson L, Nixon J. Development and validation of a new patient-reported outcome measure for patients with pressure ulcers: the PU-QOL instrument. *Health Qual Life Outcomes*. 2013;11:95.
 55. Luo Y, Yang J, Zhang Y. Development and validation of a patient-reported outcome measure for stroke patients. *Health Qual Life Outcomes*. 2015;13:53.
 56. Embretson SE, Reise S. *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum; 2000.
 57. Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, editors. *Assessing quality of life in clinical trials: methods of practice*. 2nd ed. New York: Oxford University Press; 2005. p. 53–73.
 58. Furr M, Bacharach VR. Chapter 13. Item response theory and Rasch models, *Psychometrics: an introduction*. New York: Sage; 2007.
 59. Smith LL, Reise SP. Gender differences on negative affectivity: an IRT study of differential item functioning on the multi-dimensional personality questionnaire stress reaction scale. *J Pers Soc Psychol*. 1998;75:1350–62.
 60. Meijer RR, Sijtsma K. Methodology review: evaluating person fit. *Appl Psychol Meas*. 2001;25(2):107–35.

Patient Reported Outcome Measures in Rheumatic
Diseases

El Miedany, Y. (Ed.)

2016, XI, 449 p. 84 illus., 54 illus. in color., Hardcover

ISBN: 978-3-319-32849-2