1

Introduction

1.1 What Is a PRO Measure?

The US Food and Drug Administration defines a **patient-reported outcome (PRO)** as follows: *"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. The outcome can be measured in absolute terms (e.g., severity of a symptom, sign, or state of a disease) or as a change from a previous measure. In clinical trials, a PRO instrument can be used to measure the effect of a medical intervention on one or more concepts (i.e., the thing being measured, such as a symptom or group of symptoms, effects on a particular function or group of functions, or a group of symptoms or functions shown to measure the severity of a health condition)"* (FDA 2009).

Similarly, the European Medicines Agency, in its reflection paper on the "Regulatory Guidance for the Use of Health-related Quality of Life Measures in the Evaluation of Medicinal Products," defines a patient-reported outcome as "any outcome directly evaluated by the patient and based on patient's perception of a disease and its treatment(s)" (EMA 2005). A PRO measure that will be used to collect self-reported information from patients may be as simple as a single item or as complex as a multidimensional instrument. (In this book, the terms "instrument," "measure," "questionnaire," and "scale" are used interchangeably; moreover, the terms "item" and "question" are used interchangeably.)

A PRO measure (sometimes referred to as PROM) is an umbrella term that includes a whole host of subjective outcomes such as pain, fatigue, depression, aspects of well-being (e.g., physical, functional, psychological), treatment

A Practical Approach to Quantitative Validation of Patient-Reported Outcomes: A Simulation-Based Guide Using SAS, First Edition. Andrew G. Bushmakin and Joseph C. Cappelleri.

satisfaction, health-related quality of life, and physical symptoms such as nausea and vomiting (Cappelleri et al. 2013). While the term "health-related quality of life" (HRQoL) has been frequently used instead of "patient-reported outcome," HRQoL has a broad and encompassing definition with a number of aspects or dimensions. It consumes a whole array of multidimensional health attributes collectively, including general health, physical functioning, physical symptoms and toxicity, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, and sexual functioning. Health-related quality of life can also involve toxicity and spiritual issues, among other considerations. As such, there is a looseness to the definition of HRQoL.

Health-related quality of life can be viewed as a type or category of PRO measure. Other categories of PRO measures, broadly defined, include functional status (e.g., physical function, cognitive function), which reflects an ability to perform specific activities; symptoms and symptom burden (e.g., pain, fatigue), which are specific to the type of symptom of interest; health behaviors (i.e., smoking, physical activity), which are specific to type of behavior and typically involves its frequency; and patient experience, which concerns satisfaction with healthcare delivery, treatment recommendations, and medication (or other therapies). These broad classifications are neither mutually exclusive nor exhaustive (Cella et al. 2015).

Patient-reported outcomes are, in turn, a type or category of clinical outcome assessments. A clinical outcome assessment (COA) directly or indirectly measures how patients feel or function and can be used to determine whether a treatment has demonstrated efficacy or effectiveness (Cappelleri and Spielberg 2015; FDA 2019a). There are four types of COAs: patient-reported outcomes, clinician-reported outcomes, observer-reported outcomes, and performance outcome assessments. Of the four types of COAs, the first three of these depend on the implementation, interpretation, and reporting from a rater on a patient's assessment – a rater being a patient, clinician, or non-clinician observer.

As depicted and highlighted in Figure 1.1 (FDA 2019b), **patient-reported outcome assessments**, as previously stated, rely on responses to questions on a patient's health condition that come directly from the patient, without any interpretation or judgment on the part of anyone else. An example would be a patient's self-report on a pain scale from 0 (no pain) to 10 (worst possible pain). **Clinician-reported outcome assessments** are reports of a patient's health condition from a trained healthcare professional, a member of the investigator team with appropriate clinical professional training. An example would be a clinician-reported rating scale on severity of a patient's depression (mild, moderate, severe). **Observer-reported outcomes** are reports on a patient's health condition from someone other than a patient or clinician, someone else who is not a specialized healthcare professional training, and based on observable signs, events, or

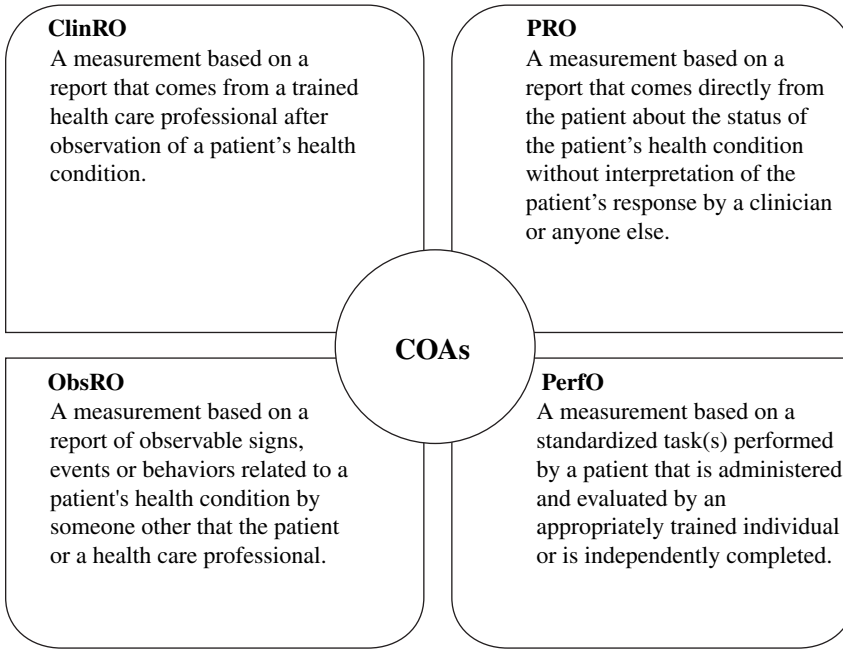

Figure 1.1 COA definitions. *Source:* Adapted from FDA (2019b).

behavior. An example would be a parent's report of infant behavior thought to be caused by pain, such as crying. The fourth type of COA, **performance outcome assessments**, involves measurements of a patient's health condition from a standard task or activity performed actively by the patient and that is administered and evaluated by a suitably trained individual or that is independently completed. An example of a performance-based activity assessment would be an exercise stress test.

Thus, a COA involves volitionally performed tasks or subjective assessments of health status that are patient-focused or centered. Among the categories of COAs, PRO measures distinguish themselves to accommodate health questionnaires whose objective is to measure a patient's self-reported health, be it HRQoL, symptom, satisfaction with treatment, functioning, well-being – whatever the purpose – from his or her perspective rather than from an alternative perspective such as a clinician, observer, or performance activity.

Patient-reported outcomes are often relevant in studying a variety of conditions – including pain, erectile dysfunction, fatigue, migraine, mental functioning, physical functioning, and depression – that cannot be assessed adequately without a

patient's evaluation and whose key questions require patient's input on the impact of a disease or a treatment. After all, who knows better than the patient him or herself? To be useful to patients and other decision makers (e.g., clinicians, researchers, regulatory agencies, reimbursement authorities), who are stakeholders in medical care, a PRO measure must undergo a validation process to confirm that it reliably and accurately measures what it is intended to measure. This validation process begins with the development of a PRO measure.

1.2 Development of a PRO Measure

1.2.1 Concept Identification

The first step in the development of a PRO measure is to identify the concept(s) to be assessed or **concept of interest** (FDA 2019b; Walton et al. 2015). The concept of interest relates to how a patient feels or functions as a meaningful aspect of health that typically occurs in a person's life. Procedures are then developed to measure the concept of interest. Concepts selected for assessment in a PRO measure should be aspects of the illness or potential benefits of treatment that are important to patients and have the potential for meaningful change, typically within the context of a clinical trial. Consider the example of self-reported pain as the meaningful health aspect to be measured. The full experience of pain has multiple conceptual faces or characteristics that include intensity, duration, frequency, and sensation (e.g., sharp, burning, stinging). Different disorders may be associated with different pain manifestations. In one disorder, for instance, the concept of interest might be pain intensity, whereas for a different disorder, it might be pain frequency.

Examples of commonly assessed concepts of interest include signs and symptoms of the disease or condition, physical functioning, psychological well-being, activities of daily living, and HRQoL. While all of these concepts may be appropriate for measurement in clinical trials in order to demonstrate treatment benefits that are important to patients, product approvals and labeling claims based on PRO data in the United States are more likely to relate to improvements in symptoms that are proximal to the disease and the drug's mechanism of action (e.g., pain, itch) than those that are more distal and complex (e.g., HRQoL). For the purpose of product labeling, concepts that are considered to be measured more objectively, such as physical function and especially signs and symptoms, also tend to be favored over more subjective concepts, such as satisfaction and self-esteem (Gnanasakthy et al. 2012, 2017). Furthermore, during the time period 2012–2016, the FDA and EMA used different evidentiary standards to assess PRO data for label claims from oncology studies: EMA was more likely than FDA to accept data from open-label studies and broad concepts such as HRQoL (Gnanasakthy et al. 2019).

Researchers recommend that the overall **context of use** for the PRO measure and its resulting endpoint(s) be considered during the concept identification step (Patrick et al. 2011a; Powers et al. 2017). The context of use describes circumstances under which the outcomes of interest adjoined to the PRO measure (or to COAs in general) have been used (i.e., labeled) or are targeted for use (i.e., they have been "qualified" or are part of an ongoing "qualification"). Such "qualification," which also applies to other COAs, is a regulatory conclusion that the PRO is a *well-defined and reliable assessment* of a specified concept of interest for use in adequate and well-controlled studies in a specified context of use (EMA 2009; FDA 2019c). PRO qualification represents a conclusion that within the stated context of use, results of the assessment can be relied upon to measure a specific concept and have a specific interpretation and application in drug development and regulatory decision-making.

Concept identification involves understanding of the disease or condition in the target population; developing an **endpoint model** for the context of use (i.e., developing a diagram of the hierarchy of relationships among the key endpoints, both PRO and non-PRO ones, that corresponds to the clinical trial's objectives, design, and data analysis plan); considering the target population (e.g., language and cultures of patients); considering preliminary issues related to instrument content and structure (e.g., the extent to which respondents will be able to identify and describe symptoms as being specific to their condition); considering the theoretical and qualitative methodologic approach (e.g., using grounded theory methods where ideas for the content of a new measure are developed inductively from the qualitative transcripts); and developing the hypothesized **conceptual framework** (i.e., developing a diagram depicting relationships between the overarching concept, hypothesized domains or sub-concepts, and candidate item content; see Section 1.2.5).

As further elaboration, an endpoint model involves listing the concepts of interest, how they are to be measured, and how the different endpoints are interrelated. For instance, Figure 1.2 illustrates an endpoint model with PRO and non-PRO assessments as key secondary endpoints and a physiologic (non-PRO) measure as the primary endpoint intended to support an indication for the treatment of Disease X (FDA 2009). In this case, the physiologic endpoint would need to show the desired efficacy in a clinical trial before success can be considered to the secondary endpoints in the trial.

1.2.1.1 Literature and Instrument Review

A review of published medical literature and other sources (e.g., practice guidelines, social media) in the relevant therapeutic area can provide a robust background for clinical aspects of the condition, symptoms, and impacts that have been already identified and existing measures that are available (FDA 2019b;

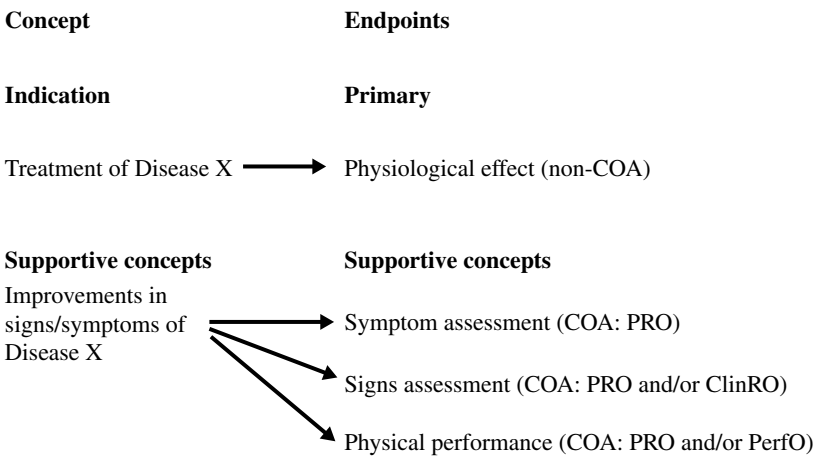

Figure 1.2 Endpoint model: treatment of disease X. *Source:* Adapted from FDA (2019b).

McLeod et al. 2018; Prinsen et al. 2016). In some cases, results of qualitative research or surveys may be available to help begin the process of identifying concepts that are important from the perspective of patients. Furthermore, the literature can provide information related to existing therapies, potential benefits of these therapies, common side effects, and the standard of care, including the role of parents or other caregivers (if relevant).

As potentially important concepts begin to emerge, instruments addressing these concepts should also be reviewed. In addition to instruments identified from previously described sources, the FDA's COA Compendium (FDA 2019d) is an excellent resource for identifying measures with the potential to support labeling claims. There are also instrument databases that can be reviewed to identify potentially relevant measures, along with information regarding the development of these measures and evidence supporting their use, such as the Patient-Reported Outcomes Measurement Information System (PROMIS®, http://www.healthmeasures.net/explore-measurement-systems/promis); the Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID, https://eprovide.mapi-trust.org/); the Measurement Instrument Database for the Social Sciences (http://www.midss.org/); and Pfizer's Patient-Centered Outcomes Assessment site (https://www.pfizerpcoa.com/).

1.2.1.2 Patient-Centered Input

The most important step in identifying concepts for measurement is the elicitation of input directly from patients (FDA 2019b). Input directly from patients is the cornerstone of content validity, which is the extent to which an instrument covers

the important concepts of the unobservable (latent) attribute or construct – that is, the concept of interest (e.g., depression, anxiety, physical functioning, self-esteem) that the instrument purports to measure. It is "the degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured" (Mokkink et al. 2010a) and thus, how well the PRO instrument captures all of the important aspects of the concept from the patient's perspective. Qualitative work with patients is therefore central to supporting content validity. While quantitative research methods are associated with the gathering, analysis, interpretation, and presentation of numerical information qualitative research methods are associated with the gathering, analysis, interpretation, and presentation of narrative information involving, for example, spoken or written accounts of experiences, observations, and events.

In the qualitative inquiry, participants are asked a series of open-ended questions aimed at furthering the researchers' understanding of the disease, identifying concepts important to patients, and observing patients' word choice when describing these concepts to inform item development (FDA 2019b; Terwee et al. 2018). Qualitative research is a scientific technique that requires a protocol outlining the study details. The protocol should include such elements as inclusion/exclusion criteria, number of subjects, pre-specification if particular sub-groups should be emphasized in recruitment, and information about the questions to be asked by developing a guide for one-on-one interviews or focus groups (or both).

One-on-one interviews or focus groups should be audio recorded to allow transcription and analysis (FDA 2019b). One-on-one interviews involve a discussion on the topic of interest between the research participant and a trained interviewer. They offer opportunities to explore topics in-depth at the individual level using probing questions. One-on-one interviewing can also be used to explore subject areas thought to be too sensitive for a focus group setting. A focus group is a carefully planned discussion conducted among a small group of participants, led by a trained moderator for the task. Group dynamics embedded in a focus group can facilitate additional insights that one-on-one interviews are not equipped for.

Consider the interview guide. It is essential to ensure consistency of questioning, and the questions should be broad and open-ended. For example, if a scale to measure sleep disorders is being proposed, examples of questions can include "How long have you had issues with your sleep?" and "What kinds of sleep difficulties do you experience?". From these broader questions, a researcher can then probe in more detail such as "You said you had problems with staying asleep, can you describe in more detail what specifically these problems are?".

Generally speaking, symptoms that patients mention spontaneously tend to be the most salient compared with those symptoms that are only endorsed in response to follow-up questions; as such, spontaneously reported symptoms may be the most important targets for measurement (McLeod et al. 2018). Additional techniques,

such as importance rating and symptom ranking, may also be used to further elucidate the relative importance of concepts elicited from patients. This same line of thinking that applies to symptoms is also relevant more generally to items of a particular concept such as mental functioning or physical functioning.

Input should be elicited from relevant stakeholders, such as caregivers, clinicians and especially patients, to inform PRO development in order to identify the core concepts to target for potential inclusion. In the development of a PRO instrument, the relevant stakeholders should be queried about important aspects of the disease or condition through one-on-one interviews, focus groups, or both – this process is referred as **concept elicitation**.

The number of patients needed for the concept identification (elicitation) stage varies depending on the amount known about the condition, the variability of disease characteristics and symptoms, the number of subgroups of interest, and the potential complexity of the PRO measure (McLeod et al. 2018). For example, the elicitation of concepts for a measure being developed to address the severity of a single symptom experienced by all (or nearly all) patients with a given disease will require a smaller sample size than a measure being developed to assess HRQoL within a diverse patient population. As a general rule, the concept elicitation process should continue until **concept saturation,** the point at which no new information is being elicited, is reached. The achievement of concept saturation should also be documented by developing a detailed table that shows the concepts elicited in each interview (or focus group) and the number of new concepts elicited in later interviews compared with earlier ones.

Once the data have been collected, analysis of the verbatim transcripts is then conducted, involving sorting quotes by concept and could be aided by using software (e.g., Atlas.ti). Such software also allows easier analysis by a specific patient characteristic (e.g., gender or disease severity) to determine if there are differences in the conceptualization of the issue under discussion.

This sorting process will lead to the development of a coding scheme whereby similar concepts are given a code name, forming a cluster. For example, if the discussion is about fatigue and how patients described this symptom, then codes might be tiredness and unrested. Direct quotes from subjects related to these specific terms would then be added under these code names, which allow the researcher to see how such terms were used and how many times the term was used. The codes are usually developed iteratively, changing as the data are analyzed, always through discussion among researchers who are generating the codes to cover the content of the transcripts from the interviews or focus groups.

This aforementioned approach is based on Grounded Theory methods (Glasser and Strauss 1999), whereby ideas for the content of a new measure are developed inductively from the qualitative transcripts. This method involves developing codes to allow the key points of the data to be gathered, which will then be grouped

into concepts, and from these concepts, a theory about the data is developed. From this process, a conceptual framework will emerge, a visual depiction of the concepts, sub-concepts, and items and how they interrelate. Other qualitative research approaches are phenomenology, ethnography, case study, discourse analysis, and traditional content analysis; a comparison of these approaches has been discussed (Lasch et al. 2010). Choice of approach will be dependent on the type of research question(s) but, for PRO development, Grounded Theory is generally preferred (Kerr et al. 2010; Lasch et al. 2010).

While patient input reigns supreme in the development of a PRO measure, input from experts in the therapeutic area or condition under study can be extremely useful throughout the development process. For example, clinical experts can facilitate concept identification by providing additional background about the condition, offering insight regarding key symptoms and impacts based on interactions with patients, and identifying unmet needs and desirable treatment attributes based on their experience.

It should be noted that qualitative data can combine and therefore complement quantitative data in collecting and analyzing the patient experience. Mixed methods research addresses a set of research questions via a hybrid approach using both qualitative (narrative) and quantitative (numerical) evidence and methods, which could be analyzed and interpreted together before a conclusion is reached. As an example of mixed methods research, a group of patients is given a survey or questionnaire with both open-ended questions (perhaps accompanied by an interview) and close-ended questions with distinct response options to quantify information for numerical analysis. Information on mixed methods research is found elsewhere (Creswell and Clark 2007; FDA 2019b; Johnson and Christen 2017; Teddlie and Tashakkori 2009).

1.2.2 Item Development

Once the concept or set of concepts to be assessed by the PRO measure has been identified, the focus of the question writing process should be on facilitating patient understanding, yielding proper response, and minimizing measurement error (FDA 2019b). Questions should be succinct and based on the language used by patients, maintain an accessible reading level, and naturally relate to the response options. Items should be comprehensible and generalizable to the total target population, independent of education level, with difficult words and complex sentences avoided, such that anyone over 12 years of age can understand them (Streiner et al. 2015). Item developers should be careful to avoid asking two or more questions masquerading as a single question that addresses more than one concept, which might have different or conflicting answers. For example, an item may ask if a person has trouble walking or jogging. When answering this

item, an individual who can walk but not jog must determine which component to answer and which to ignore (McLeod et al. 2018).

Response options should be both comprehensive and mutually exclusive; it is also important that patients be able to clearly differentiate among each option. Responses with multiple meanings should be avoided. For instance, the word "fair" can mean "pretty good," "honest," "plain," and "according to the rules" (de Vet et al. 2011). Careful attention must also be given to the recall period to ensure that patients are able to accurately remember relevant information and formulate an accurate response without complex computations. Thus, it should be clear to which period of time the question refers. Should the patient respond to current pain, pain during the past 24 hours, or pain during the previous 7 days? While very short recall periods (e.g., past 24 hours) and frequent administration have been recommended by the FDA for assessing symptoms with the potential to vary from day to day, as well as within a day, there may be situations in which a longer recall period and less frequent administration are more appropriate. Lengthier recall periods may be appropriate for the assessment of symptoms, behaviors, and functions that are slow to change or require an extended period to assess accurately, such as, for example, using four weeks to assess erectile functioning.

In the context of clinical trials, measures need to be designed to evaluate treatment benefits. As such, patients' experiences (and, consequently, their responses to the PRO items) must have the potential to change over the time course of a clinical trial. Patients typically make multiple measurements over time across multiple PRO instruments. Therefore, the PRO measure should be brief in order to minimize patient burden and allow for flexibility in the mode of administration. For example, a PRO measure being developed for electronic administration on a daily basis should not include more items than patients are willing to answer consistently; the individual items should be brief enough to display clearly on a small screen (e.g., hand-held device or smartphone).

Multiple items may be drafted for each concept of interest so that variable question wording and response scales can be tested in cognitive debriefing interviews and psychometric evaluation to identify what version of the item or its response scale (or both) is most easily understood and answered by patients. In the measurement of pain, for instance, it is not expected that patients can reliably discriminate between 65 and 67 mm of pain or even between 65 and 67 mm on a continuous (interval) 100-mm line visual analog scale (0 = no pain to 100 = worst imaginable pain). Instead, a numeric rating scale with 11 discrete (ordinal) categories would be preferred (0 = no pain, 10 = worst imaginable pain). Of related concern is to set the number of categories on an ordinal scale based on how many categories are relevant.

In addition, it is helpful to consider potential scoring rules during item development. Consider the item "Rate your current level of pain." Including a category

for "no pain" can be justified as the lowest (most favorable) score for patients who do not have pain; however, if a "not applicable" response option is offered instead, this response may be chosen by patients who do not experience pain or by patients who restricted their activity to avoid pain, creating measurement error in the item score.

This subsection provided a few examples of requirements in formulating adequate items and responses. A more detailed exposition on the essentials for adequate formulation of questions and answers can be found elsewhere based on principles used in survey research methodology (Bradburn et al. 2004; Fowler 1995; Presser et al. 2004; Schwarz and Sudman 1996; Stone et al. 2000; Sudman et al. 1996).

1.2.3 Cognitive Interviews

Cognitive interviews involve incorporating follow-up questions in a field test interview to gain a better understanding of how patients interpret questions asked of them. Cognitive interviews are a qualitative research tool used to determine whether concepts and items are understood by patients in the same way that instrument developers intend. In this method, respondents are often asked to *think aloud* and describe their thought processes as they answer the instrument questions (FDA 2019b). In general, cognitive interviews are debriefing sessions that focus on the cognitive processing involved when participants read, interpret, and determine their responses to the draft questions.

Insights gained through evaluating this processing step are then used to refine instructions, question-wording, response categories, formatting, and other aspects of the PRO measure to remove aspects that are unclear or influence participants to interpret the items in a way that was not intended; the process is iterative and should continue until evidence is established that no further revisions are warranted. Cognitive interviews typically involve the collection of supplemental data to further support content validity and inform future use of the final measure. As with the concept elicitation interviews or focus groups, cognitive interviews should be accompanied by a semi-structured interview guide; doing so will ensure consistency among cognitive interviews. Additional information on the conduct and analysis of data collected during cognitive interviews can be found in multiple sources and the references cited therein (FDA 2019b; Willis 2005, 2015).

The goals of cognitive interviews are essentially a twofold process: (i) to ensure that the most important concepts are included in the final PRO measure; and (ii) to ensure that respondents understand how to answer each item based on clear instructions, the appropriate recall period, item meaning, response scale, and any other features, such as paper versus electronic mode of administration (Patrick et al. 2011b; Powers et al. 2017). In addition, if subscales are desired for

specific concepts of interest, gaining an understanding of how patients perceive relationships among the items will help inform the structure of the PRO scale and the development of initial scoring algorithms. It is also useful to pose specific questions to ascertain what amount of change on the individual items and sets of items is meaningful to patients in order to facilitate the development of Meaningful Within-Patient Change definitions for applications in future clinical trials.

As with the concept elicitation step, cognitive interviews will vary in the number of patients needed as a function of the complexity of the concepts to be measured and variability across the patient population. In general, however, three or four iterative sets of interviews, each comprised of 6–10 patients (for a total of 18–40 additional patients), are typically sufficient to help refine and finalize the items. As a general rule, the cognitive interviews are complete when no additional item changes are identified by subsequent interviews (McLeod et al. 2018).

An **item tracking matrix** is a record of the development (e.g., additions, deletions, modifications, and reasons for the changes) of items or tasks used in an instrument. The item tracking matrix is commonly requested by FDA reviewers and describes the item wording for each version of each item (question) tested, reasons for revisions to retained items, and reasons for the omission of items during the process (FDA 2019b). Moreover, the item tracking matrix provides evidence that further refinements are not needed by documenting subsequent interviews that attest to sufficient understanding, with no feedback identifying weaknesses or suggesting revisions to clarify the wording of the questions or response choices.

1.2.4 Additional Considerations

It should be noted that the steps outlined in the preceding sections are intended to be a general guideline and not prescriptive. As such, suppleness is required when developing a PRO measurement. While not intended to provide an exhaustive or a complete list of special circumstances, this subsection highlights some additional considerations.

For example, if a PRO measure is meant to address a single concept (such as the severity of a particular symptom), or if a great deal is already known about the concepts that need to be measured based on prior research, it may be reasonable to combine the concept identification and cognitive interviewing steps. Interviews would begin with a comprehensive concept elicitation phase in addition to refining the existing item set in supporting the content validity of the final measurement. Such an approach may also be appropriate if interest centers on modifying an existing PRO measure rather than developing a new measure.

The development process often needs to be tailored when working with special populations, such as children and individuals with cognitive or sensory

impairments. Recommendations for the conduct of qualitative research in such patient populations have been provided (DeMuro et al. 2012).

If a PRO measure is meant to be administered in international trials, the development process should ideally be completed in multiple countries. It is very common, however, for the initial development of a PRO measure to be limited to a single country. In such cases, translation and cultural adaptation are then performed later in preparation for administration in international studies.

When instrument development takes place in a single country, but the measure is expected to be used in international trials, it is often useful to conduct a translatability assessment. This assessment includes an evaluation of the draft items by translation experts to identify any potential issues with the item wording (language or concepts) and response options/scales. Translatability assessments should be conducted during the item refinement process so that modifications suggested by translators can be tested with participants in additional debriefings from cognitive interviews. For these activities, methodology has been recommended (Wild et al. 2005).

Considerations related to the development of a PRO measure for electronic administration (ePRO) are analogous to those pertaining to its development in multiple languages (McLeod et al. 2018). Ideally, the development process involves cognitive interviewing of ePRO versions of the items with patients using the device slated for use in clinical trials. This process negates the need to demonstrate measurement equivalence between different modes of administration (e.g., ePRO with paper), as the ePRO version is the original and only version of the measure. Nonetheless, the development of a paper-based PRO measure that is later migrated to an electronic platform or migrated from one type of electronic platform to another (e.g., web-based to hand-held) is not uncommon. Much like translation and cultural adaptation processes, ePRO migration must follow a rigorous process to ensure that the content validity and psychometric properties of the final version are maintained. The evidence required to support measurement equivalence between paper-based and ePRO versions has been described elsewhere (Coons et al. 2009).

1.2.5 Documentation of Development Process with Conceptual Framework

It is important to document all steps involved in the PRO development process from the literature review, expert feedback, qualitative research and cognitive interviews for concept elicitation and modification, item development and refinement, endpoint model diagram, and, eventually, current concept and instrument selection up until that point in the process. In addition, from this multi-staged process, a **conceptual framework** for the new measure should be depicted and described. The conceptual framework of a PRO instrument is an explicit description or a diagram

for an instrument showing the relationship among items (i.e., questions or tasks) included therein, domains (domain-level concepts), and concepts measured, along with the scores produced by the instrument (FDA 2009, 2019b).

As instrument development evolves, the conceptual framework evolves in tandem and its concepts measured by domain or total scores should be derived using patient input to ensure that the conclusions drawn using instrument scores are valid. Of interest is how individual items are thought to be related, how items may be related within a domain, how multiple domains may be related to each other, and the general concept of interest as reflected in the conceptual framework of the PRO instrument. Figure 1.3 depicts a generic example of a conceptual framework where Domain 1, Domain 2, and Overall Concept each represent related but separate concepts of interest. Conceptual frameworks apply generally to the four types of COAs, not just PRO instruments.

Related to but distinct from a conceptual framework is a **measurement model**, which examines the relationship between the latent variables (domains) and their observed items or variables or questions **(indicators**). In a **reflective model,** a latent variable drives the indicators (known as **effect or reflective indicators**), which involves a majority of measurement work in the social and health sciences.

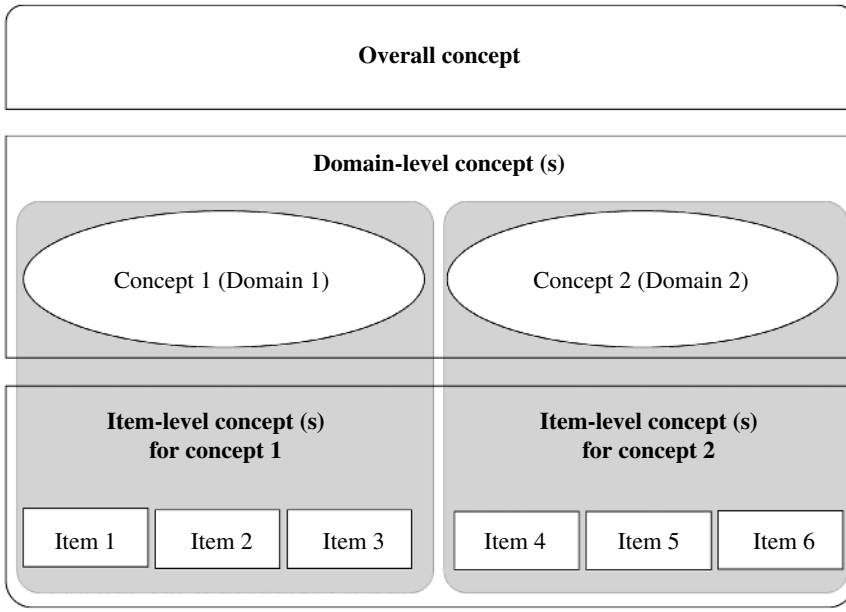

Figure 1.3 Diagram of the conceptual framework of a PRO instrument. *Source:* Adapted from FDA (2019b).

Here, the latent variable drives its effect indicators; in other words, all items should change in the expected way when the concept or construct changes (i.e., items should **reflect** the changes in the concept – hence the name of the model). For instance, as the construct of anxiety changes, so do the items that measure it such as tension, fears, and insomnia.

In a **formative model,** on the other hand, the indicators (also called **causal or formative indicators**) influence its concept; therefore, a change in an indicator affects a change in the concept or construct (rather than the reverse). At the same time, not all causal indicators may change to affect the construct – a change in one of the items may suffice to change the concept, and not all items need to be related. Side effects are an example of variables that are causal indicators in relation to a construct of overall health-related quality of life. More detail on reflective models and formative models are found elsewhere (de Vet et al. 2011; Fayers and Machin 2016) and will also be discussed in detail in Chapter 5 (specifically in Section 5.2.5).

The conceptual framework could be as simple as a single item addressing the severity of a single symptom such as pain intensity or as complex as groupings of items underlying various domains of HRQoL. If the concept of interest is a general one like physical function or clusters of specific symptoms and signs, which includes multiple items and domains, a single-item PRO endpoint will be inadequate to support labeling claims about each of the different aspects of the general concept.

The conceptual framework provides an initial glimpse into possible scoring for the PRO measure based on the development phase (e.g., which items belong to which domain in forming a domain-specific score, and which domains are combined in forming a total score) and may be refined based on the results of the quantitative (psychometric) phase in which the relationships among item scores are evaluated empirically. Documentation of the development process, including the conceptual framework, should be part of the regulatory dossier to facilitate review and agreement that the PRO measure under consideration is fit for purpose – that it is a suitable measure in the proposed context of use for supporting medical product labeling.

1.3 Psychometric Validation

Several guidance on psychometric validation of PRO measures are available from regulatory documents (EMA 2005; FDA 2009, 2019b), books (Bartolucci et al. 2016; Cappelleri et al. 2013; de Vet et al. 2011; DeVellis 2017; Fayers and Machin 2016; Mair 2018; Streiner et al. 2015; Walters 2009), and articles (Calvert et al. 2013, 2018, 2021; Chassany et al. 2002; Mokkink et al. 2010a, 2010b;

Reeve et al. 2013; Revicki et al. 2007a). They center around essentially the evaluation of reliability, validity, and ability to detect change (often referred to as responsiveness) as the psychometric groundwork for the PRO measure as part of good measurement science, especially to support product and labeling claims or, simply, to conduct good measurement science. **Reliability** is the extent to which the measure yields reproducible and consistent results (when it should), **validity** is the extent to which an instrument measures what it is intended to measure and can be useful for its intended purpose, and **ability to detect change** is the ability of the measure to identify differences in scores over time in groups or individuals who have changed with respect to the measurement concept. In general, the quantitative assessment of assessing these measurement properties should be made within the same context of use as planned for the pivotal clinical trials to support labeling claims (if the intention is a label claim).

Generally, the same set of measurement properties should be evaluated regardless of the intent of the PRO measure or whether the PRO measure is newly developed or an existing measure being used in a different condition or for a different purpose. For a PRO instrument that is intended to support product approval or labeling, the amount of evidence required is substantial in order to ensure that the measurements of concepts important to patients are reliable, valid, and have the ability to detect change, as well as be capable of detecting a known treatment benefit.

1.3.1 Psychometric Evaluation Data

Preliminary psychometric evaluations are suggested using a non-interventional study, which is typically based on cross-sectional data, outside of the clinical trial program. In addition, a comprehensive psychometric evaluation of a PRO instrument, including using longitudinal data, should be conducted during a Phase 2 randomized controlled trial. It is important for the Phase 2 study, as well as a non-interventional study, to include additional measures that can support the psychometric evaluation: measures of similar constructs, global assessment of disease status, and global change in disease status relative to baseline. Whenever possible, given the subjectivity of PRO measures, a trial with complete blinding is the first choice.

The PRO measure evaluated before Phase 3 should be administered at key time points aligned to the intended time points for future pivotal randomized controlled trials in Phase 3, which can provide supplemental and confirmatory psychometric information based on previous studies. In general, it is too much of a risk to rely on pivotal Phase 3 studies to establish the major psychometric characteristics of a PRO measure (unless the intention is to seek a label claim on a PRO measure from a Phase 4 study as may occur, for instance, through a supplemental new drug application). In Phase 2, these time points should provide

opportunities to evaluate change in the proposed scores where change is anticipated (assuming a known relative treatment effect from the experimental intervention compared with the control intervention) and to evaluate the stability of scores where change is not expected (during the pre-treatment stage, before randomization).

Sensitivity, the ability of a measure to detect differences between groups of patients, is important in clinical trials as a PRO instrument that cannot differences in patient outcomes between randomized groups would have little value. Selection of the number of additional measures should be prudent and not increase patient burden unnecessarily. Furthermore, it is important that the selected study has adequate sample size for the methods planned (Chen et al. 2014).

1.3.2 Psychometric Properties

The psychometric evaluation should be guided by a formal and prespecified psychometric analysis plan. If the psychometric evaluation is performed using clinical trial data (e.g., Phase 2 data), the psychometric analysis plan on the PRO instrument should be a separate document from the main statistical analysis plan designed to guide the evaluation of efficacy and safety for the clinical trial. If the PRO measure or measures are part of the multiple testing procedures of the efficacy outcomes (both PRO and non-PRO) implemented to preserve type I error (i.e., the chance of incorrectly or falsely declaring a statistically significant result), such detail should be part of the main statistical analysis plan.

The psychometric statistical analysis plan should describe the psychometric and statistical methods to be conducted, including prespecified analytic samples and subgroups of interest. The psychometric (statistical) analysis plan should provide a detailed description of analyses to assess reliability, validity, and ability to detect change and, in interventional studies, thresholds for the clinically meaningful between-patient difference and especially clinically meaningful within-patient change. It should be noted that problems with content validity (Section 1.2) cannot be rectified by testing other measurement properties; content validity is an indispensable prerequisite for a proper quantitative evaluation.

While content validity is of particular interest during the qualitative phase, **construct validity** is of particular interest during the quantitative stage. Construct validity consists of evidence that relationships among items, domains, and concepts conform to a priori hypotheses concerning logical relationships that should exist with other measures or characteristics of patients and patient groups (FDA 2009). Generally, the statistical analyses at baseline and post-baseline (follow-up) would combine all treatment groups as one unit rather than having analysis for each treatment group. Table 1.1 provides an overview of typical analysis methods for each of the key psychometric properties.

Table 1.1 Key psychometric properties assessed.

Psychometric property	What is assessed?	Evaluation method
Distributional characteristics	Descriptive assessment of the target PRO measure on item-level and scale-level scores at key time points	Mean, median, standard deviation, minimum, maximum, percentage missing, frequency distribution
Measurement model structure	To assess if the measurement model of a multi-item scale fits the data	Confirmatory factor analysis
Reliability		
Internal Consistency	Consistency of responses to items on the same multi-item scale, where the items are intended to tap into different aspects of the same underlying concept	Cronbach's coefficient alpha
Test-retest	Stability of scores over time when no change is expected in the concept of interest	Intraclass correlation coefficient
Validity		
Content	Evidence that the instrument measures the concept of interest including evidence from qualitative studies that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use	Qualitative evaluation of stakeholders (especially patients) involving interviews or focus groups and cognitive interviewing to evaluate patient understanding
Construct	Evidence that relationships among items, domains, and concepts conform to a priori hypotheses concerning logical relationships expected to exist with similar or dissimilar measures Includes at least two major elements: 1) strength of correlation testing a priori hypotheses (convergent validity with similar measures and divergent or discriminant validity with dissimilar measures) and 2) degree to which the PRO instrument can distinguish between or among groups hypothesized a priori to be different (known-groups validity)	Correlational analyses, anchor-based analyses, effect size statistics

Table 1.1 (Continued)

Psychometric property	What is assessed?	Evaluation method
Ability to detect change	Evidence that a PRO instrument can identify changes in scores over time in individuals or groups who have changed with respect to the measurement of concept	Anchor-based analyses complemented with correlational analyses, effect size statistics
Meaningful change and difference	Thresholds for meaningful within-patient change and, separately, between-group difference on the target PRO measure	Anchor-based analyses, cumulative distribution function plots, probability density functions, effect size statistics

Note: PRO, Patient-reported outcome.

1.3.2.1 Distributional Characteristics

Key demographic and sample characteristics at baseline should be tabulated to describe the sample used in the psychometric evaluation. Standard descriptive statistics (including mean, median, standard deviation, minimum, maximum, and percentage missing) should be reported at baseline and, for an interventional study, key post-baseline time points for scores on the targeted PRO measure.

A floor effect is defined by a high proportion of subjects selecting the minimum category or score (e.g., more than 15% select the minimum possible score for scales with a wide range of scores like the SF-36 [McHorney and Tarlov 1995; Terwee et al. 2007]). Alternatively, a ceiling effect is defined by a high proportion of subjects selecting the maximum possible score (e.g., more than 15% select the maximum possible score for scales with a wide range of scores like the SF-36 [McHorney and Tarlov 1995; Terwee et al. 2007]). For reflective indicators (items), both floor effects and ceiling effects can indicate that the PRO measure may not have the measurement range necessary to differentiate and show change. For instance, because patients tend to be homogeneous in their responses at baseline, before treatment intervention, such responses across patients typically have less covariation with a restricted range of scores stemming from a circumscribed inclusion and exclusion criteria. For formative indicators, such as symptoms, a floor or ceiling effect on an item may still be extremely important for descriptive or evaluative purposes (Fayers and Machin 2016).

Given that the magnitude of correlations between variables is a function of their variability (all other factors equal, a restricted range or less variability yields

a smaller or attenuated correlation in absolute value), baseline is usually not the ideal time for psychometric evaluation. Under the assumption of a beneficial treatment effect, the range of data expands at post-baseline and the larger estimate of the correlation coefficient (in absolute value) is more accurate reflection of the true correlation coefficient (Bland and Altman 2011; Crocker and Algina 1986; Nunnally and Bernstein 1994).

Thus, if a majority of subjects tend to select an extreme category (or categories) for pain severity at baseline, as a result of the inclusion and exclusion criteria, the estimated correlation coefficient between a targeted PRO pain measure and a related supportive measure would be expected to be lower and attenuated relative to what it would be during the treatment phase when the responses are expected to be more diversified and varied owing to treatment benefit. A careful review of the concept addressed by the item and the characteristics of subjects who endorse extreme categories should be examined to more accurately determine whether an item is only applicable to a certain segment of patients in the target population or simply an artifact of the sample owing to the circumscribed inclusion and exclusion criteria. Therefore, exploration of potential reasons for the observed floor effect or ceiling effect will help to inform decisions on removing, revising, or retaining an item under consideration.

Review of the distributional characteristics provides a general evaluation of compliance and suitability of the PRO measure. Evidence of good compliance (i.e., minimal missing data across items and time points) provides support that patients are capable of completing the measure and that the patient burden of completion is acceptable. The appropriateness of scale-level and item-level scores are judged by the distribution of response categories across time points. Supportive evidence is defined by distributions where all categories are selected by at least a small proportion of subjects without evidence of floor and ceiling effects.

Evidence that the distributions for the PRO instrument and its supporting measures behave as expected provide an initial quantitative understanding for the target PRO measure. For example, if there is little variability in both the target PRO measure and in a related supportive measure, then the concern may shift to the relevance of the sample or overall context of use. On the other hand, if the supporting measures have the expected distributions but the target PRO measure does not – especially during post-baseline assessments with a beneficial intervention – the initial descriptive statistics for the target PRO measure may foreshadow problems with other measurement properties (McLeod et al. 2018).

1.3.2.2 Measurement Model Structure

For multiple items with the potential for subscales, the conceptual framework provides a description of the item-level relationship based on the qualitative phase of the development process and how the items form potential domains.

To inform scoring, this framework is further refined by evaluating the quantitative relationships among the items through inter-item correlation coefficients and dimensionality analysis (e.g., factor analysis).

Methods such as **exploratory factor analysis** have been recommended when these frameworks are preliminary and alternative item groupings have been considered (Finch 2020). In exploratory factor analysis, there is initial uncertainty as to the number of factors (latent variables or domains) being measured, as well as which items are representing those factors. Exploratory factor analysis is suitable for generating hypotheses about the structure of distinct concepts and which items represent a particular concept. Moreover, exploratory factor analysis can further refine an instrument by revealing what items may be dropped from the questionnaire, because they contribute little to the presumed underlying factors.

That said, however, we recommend that, for modern development of PRO instruments, we should already know from the qualitative research and conceptual framework which items belong to which domain. Thus, based on the qualitative research and conceptual framework, we should perform confirmatory factor analysis rather than exploratory factor analysis. Confirmatory factor analysis facilitates the evaluation of a proposed structure (Brown 2015). In this type of analysis, the number of factors and the items that should load on each factor are prespecified. Items are grouped based on the conceptual framework to form hypothesized domains. In confirmatory factor analysis, factor loadings are coefficients that indicate the strength of the relationship between an item and the latent factor (or how well an item reflects changes in the latent factor). These coefficients are important because they signify the nature of the variables that most strongly relate to a factor; the nature of the variables helps to capture the nature and meaning of a factor.

Confirmatory factor analysis is a hypothesis-confirming technique relying on a researcher's hypothesis that requires pre-specification of all aspects of the measurement model: the number of factors, the pattern of item-factor relationships, and so forth. Confirmatory factor analysis tests whether the pre-specification measurement model fits the data.

It is important to stress that baseline data from a clinical trial is not the best or only choice to validate a PRO measure due to the restricted range expected in the data at baseline, as noted previously. Having a wider net of responses (owing to a presumed treatment benefit), post-treatment visits should be especially considered and, relative to baseline, are expected to be more representative of the underlying covariation and, as such, better suited to assess psychometric properties of the scale.

In addition, Rasch modeling and other types of item response theory modeling are useful tools to evaluate the performance of individual items (Andrich 1988; Cappelleri et al. 2014; Edelen and Reeve 2007). These methods can be used to

assess the relationships among items, how the items relate to an underlying construct, and redundancy in item content, making the methods informative for item reduction and subscale (domain-level) refinement.

1.3.2.3 Reliability

Reliability is defined as the extent to which an instrument is free of measurement error and can consistently measure a subject's true score, which is the average score that would be expected if the subject completed parallel forms of the instrument many times (Hays and Revicki 2005). A perfectly reliable scale would be a reflection of the true score and nothing else. It is important to evaluate the reliability of a PRO measure in order to provide evidence for the reproducibility of its reported scores under stable conditions. For PRO measures, internal consistency reliability and test-retest reliability are the most relevant forms of reliability.

Internal consistency reliability applies to the consistency of responses to items on the same multi-item scale, where the items are intended to tap into the same construct (or complementary interrelated aspects of it). This form of reliability uses the correlation and covariance among items on a multi-item PRO scale to assess the homogeneity (similarity) of its items at a particular time and refers to the extent to which the items are interrelated and covary. Internal consistency reliability is typically estimated for a group of items whose responses sum to form a single score (a domain score) using Cronbach's coefficient alpha (Cronbach 1951; DeVellis 2017).

Cronbach's alpha presumes that all items adjoined to a multi-item PRO scale reflect a single (latent) concept or construct of interest and is therefore unidimensional, such as scales on anxiety, cognitive ability, and physical function. Thus, Cronbach's alpha coefficient is intended for a PRO measure whose items are intended to summarize distinct aspects of a condition or disease that are related (e.g., different items for measuring anxiety) and not for a PRO measure whose items may not be related (e.g., differing symptoms that are indicative of a disease or condition but may not typically present together).

Test-retest reliability provides an evaluation of reliability by comparing scores for subjects who are classified as stable on the measured construct across at least two time periods where no change is anticipated. The PRO scale is measured at one time (test) and then at least at one other time (retest). Selecting the right period between test and retest, or between times or occasions, is crucial. Typically, a stable group predefined naturally as exhibiting no change during the pretreatment period in clinical trials or may be based on a criterion measure that is judged to be able to adequately assess patients' status in clinical studies.

An intraclass correlation coefficient (ICC), is then computed using the scores for the stable subgroup at specified time points in order to quantify test-retest reliability (McGraw and Wong 1996; Qin et al. 2019; Schuck 2004). An ICC is the ratio

of the between-subject variability to total variability. If this ratio is high, there is little measurement error (within-subject variability) to be accounted for by the measurements on the different occasions and the reproducibility (i.e., test-retest reliability) of the measure is high. Values of ICC of at least 0.70 for multi-item scales are recommended to support adequate test-retest (Nunnally and Bernstein 1994; Reeve et al. 2013).

1.3.2.4 Construct Validity

Construct validity is defined as the degree to which the scores of a measurement instrument are consistent with hypotheses, for example, with regard to internal relationships, relationships with scores of other instruments, or differences between relevant groups (de Vet et al. 2011; Mokkink 2010a). Without adequate evidence to support content validity (Section 1.2), most regulatory reviewers will dismiss a measure for further review of the other aspects of validity. These additional aspects of validity use quantitative data and methods grounded in construct validity to build upon the qualitative evidence and prior or preliminary quantitative evidence to further support the content validity of the PRO instrument. Two major types of construct validity are (i) convergent and divergent validity and (ii) known-groups validity.

Convergent/divergent validity describes the prespecified relationships among multiple indicators (items or questions) of constructs (concepts of interest) and the degree to which the scores from these indicators follow predictable patterns. The goal of these analyses is to demonstrate stronger relationships among measures addressing similar constructs (defined as "convergent" validity) compared with measures addressing more disparate or dissimilar constructs (defined as "divergent" or "discriminant" validity). Typically, correlational analyses are conducted to examine these relationships using data at baseline and other key time points. Specifically, the psychometric analysis plan should outline a priori hypotheses that specify the anticipated direction and magnitude or strength of correlations between scores on the target PRO measure and scores on other supporting measures included in the trial (i.e., other PRO measures of both similar and different constructs, other types of clinical outcome assessments, or both). Correlation analyses are then conducted, and their magnitudes and patterns are compared against a priori hypotheses. In most circumstances, a correlation larger than 0.4 is reasonable to support evidence for convergent validity, less than 0.3 to support divergent validity, and between 0.3 and 0.4 as inconclusive evidence to support convergent or divergent validity.

Known-groups validity, another form of construct validity, is based on the principle that the PRO measurement scale of interest should be sensitive to differences between specific groups of patients known to be different in a relevant way. For instance, if a PRO instrument is intended to measure visual functioning,

its mean PRO scores should be able to differentiate sufficiently between subjects with different levels of visual acuity (mild, moderate, and severe). Again, these assessments begin with a priori hypotheses that are then evaluated empirically in terms of their predicted direction and magnitude. The magnitude of the separation between known distinct groups is more important than whether the separation is statistically significant, especially in studies with small or modest sample sizes where statistical significance may not be achieved.

While known-group validity and convergent/divergent validity are among the most common forms of construct validity (Table 1.1), other forms of construct validity also have their relevance (Cappelleri et al. 2013; de Vet et al. 2011; Mokkink et al. 2010a). One of them is **structural validity (**i.e., **measurement model structure)**: the degree to which the items of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured, which is assessed by confirmatory factor analysis.

Cross-cultural validity, another type of construct validity, involves the extent that the performance of items on a translated or culturally adapted PRO instrument is an adequate reflection of items in the original version, which is often assessed after the translation of the questionnaire.

Another form of validity, which is sometimes grouped with construct validity but more often is left distinct, is criterion validity, with its two types being **concurrent validity** and **predictive validity**. **Criterion validity** involves assessing a PRO measurement instrument against the true value or against another standard indicative of the true value of measurement, which provides an adequate reflection of a gold standard: **concurrent validity** involves an assessment of scores from the targeted PRO measure with the scores from the gold standard (or quasi-gold standard) at the same time, while **predictive validity** involves an assessment of how the targeted PRO measures predict the gold standard (or purported gold standard) in the future.

1.3.2.5 Ability to Detect Change

The **ability to detect change** – or **responsiveness** – refers to evidence that a PRO instrument, as with other COA instruments, can identify differences in scores over time in individuals or groups who have changed with respect to the measurement concept (FDA 2009, 2019b). In its draft guidance (as of this writing), the FDA reviews the ability to detect change on an instrument as follows: "*FDA reviews a COA's ability to detect change using data that compares change in COA scores to change in other similar measures that indicate that the patient's state has changed with respect to the concept of interest. A review of the ability to detect change includes evidence that the instrument is sensitive to gains and losses in the measurement concept and to change across the entire range expected for the target patient population. When patient experience of a concept changes, the value(s) for the COA measuring that concept also should change*" (FDA 2019b).

Therefore, one way to assess the ability to detect change is to map change from baseline in the target PRO measure with change from baseline on another measure known as an **anchor measure**, an external measure (which can be another PRO measure or other COA measure) that should be clearly interpretable and appreciably correlated with the target PRO measure (FDA 2019b; King 2011). An example of an anchor measure includes the patient global impression of severity – for instance, "Please choose the response that best describes the severity of your overall status over the past week: none, mild, moderate, severe, very severe." Another example of an anchor measure is patient global impression of change – for instance, "Please choose the response that best describes the overall change in your overall status since you started taking the study medication: much better, a little better, no change, a little worse, much worse." Of these two types of measures, the former one which is based on a static and current global impression of severity is recommended at minimum. It is less likely to be subject to recall error than the latter one, which is based on retrospective global measure of change, which can be used as secondary anchor in sensitivity analysis.

1.3.2.6 Interpretation

Unlike well-established clinical measurements such as survival and blood pressure, which are generally understood and can be measured directly, the latent or unobserved concepts captured by PRO measures may be unfamiliar to many healthcare professionals and patients. Patient-reported measures may lack sufficient data or experience or clinical understanding to draw from to properly interpret what, for example, is a 10-point change means on a 0–100 PRO scale.

While PRO endpoints in clinical trials are commonly assessed through the statistical comparison of group means, the results of these comparisons are not always easy to interpret. In addition, statistical significance alone is not sufficient to demonstrate clinically important benefits. As such, characterizing meaningful change for individual patients and meaningful difference at a group level provides the ability to further evaluate, interpret, and communicate PRO results to regulators, patients, and prescribers.

Individual within-patient change is different from between-group mean difference or treatment effect. From a regulatory standpoint, FDA is more interested in what constitutes a meaningful within-patient change in scores from the patient perspective (i.e., at the individual patient level). The between-group mean difference is the difference between the average change score (or average score) between two study arms that is commonly used to evaluate treatment difference, but it does not address the individual within-patient change that is used to evaluate whether a meaningful score change is observed; therefore, a between-group treatment effect is different from a meaningful within-patient change (FDA 2019b).

The literature related to interpretation of change and difference on PRO measures is vast and many terms have been used to characterize them, often interchangeably. For individuals, the terms "responder definition," "responder threshold," and "clinically important change" are among those used; for groups, the terms "minimal important difference," "minimal clinically important difference," and "clinically important difference" are among those used (Cappelleri and Bushmakin 2014; Cook et al. 2015; Coon and Cook 2018; Coon and Cappelleri 2016; Crosby et al. 2003, 2004; FDA 2009, 2019b; Giesinger et al. 2020; King 2011; Marquis et al. 2004; Revicki et al. 2007b, 2008). Given the variety of methods and labels, it is essential to define, understand, and formulate the level of change or difference unambiguously.

1.4 Learning Through Simulations

This book is intended for researchers in the biopharmaceutical and healthcare industry, government, and academe involved in health measurement scales, especially PRO measurement and analysis. Intended readers of the book include statisticians, statistical programmers, psychometricians, and other researchers such as epidemiologists and those involved in outcomes research.

The ensuing chapters provide a detailed description of key steps needed for quantitative validation of PRO measures. Exposition covers major aspects of psychometric validation methodology on PRO measures, including exploratory and confirmatory factor analysis, reliability assessments, longitudinal analysis, and meaningful changes and differences on PRO measures.

This simulation-based guide is intended to give the reader an in-depth understanding on how initial simulated values, various data patterns, and different assumptions affect the results. Much comprehension can be gained by applying computer simulation methods to teach statistics (Good 2005; Mair 2018; Morris et al. 2019; Wicklin 2013) including psychometrics (Beaujean 2018; Feinberg and Rubright 2016). In the chapters that follow, common themes are covered such as how to structure the data before PRO analysis, how to implement an analysis in a step-by-step fashion using simulation examples, and how to delve deeply into concepts by modifying assumptions and parameters.

In order to accomplish these topics of instruction, the book centers on providing a deep understanding and enriched learning environment of quantitative validation methods for PRO measures via simulated data and analytic implementations using the SAS® software. Given that the biopharmaceutical industry is the major sponsor of PRO measures – and that the SAS software is the standard software used in this industry (including by regulatory agencies) – the use of the SAS®

software is a natural choice. In academic settings, faculty members and students generally have free access to SAS software. And, in August 2021, the SAS Institute created a cloud-based version that is available for everyone who would like to learn, even though it is still titled as "SAS OnDemand for Academics": https://www.sas.com/en_us/software/on-demand-for-academics.html. An independent learner simply will need to create a free account with the huge advantage that nothing is needed to be installed on your computer, as everything will be done in the cloud. All that is needed is a compatible web browser. The website https://www.sas.com/en_us/software/on-demand-for-academics.html also provides a lot of additional resources to learn SAS and SAS programming, such as, for example, 10-minute video "Getting Started With SAS Studio" or the free e-learning course "SAS Programming 1: Essentials."

1.5 Summary

This chapter is intended to be general in highlighting the definition and context of PRO measures. The development of a PRO measure is discussed and grounded in concept identification, item development, cognitive interviews, and other considerations. The chapters provide a general view of the instrument development process through the conceptual framework.

After the crucial and indispensable role of qualitative research is described as the basis content validity, the chapter covers psychometric validation of a PRO measure by describing the type of studies and data needed; the measurement properties investigated in terms of reliability, construct validity, and ability to detect change; and methods for interpretation of meaningful scores. Broad-stroke elements in an SAP are noted. The role of simulation-based learning using the SAS software for understanding psychometrics is emphasized. Subsequent chapters cover psychometric themes that are conceptualized and explained and then illustrated and augmented through simulation-based learning activities.

In what follows, this book assumes that the PRO instrument of interest has achieved (or is in the process of achieving) content validity, chiefly through qualitative research, in that it covers the important concepts of the unobservable or latent attribute under study (e.g., depression, anxiety, physical functioning, self-esteem) that the instrument purports to measure. Hence the content of the PRO instrument is taken as an adequate reflection of the construct to be measured. Given this backdrop, subsequent chapters focus on the quantitative validation of PRO measures in particular and, when applicable, to COAs in general.

References

Andrich, D. (1988). *Rasch Models for Measurement*. Newbury Park, CA: SAGE Publications, Inc.

Bartolucci, F., Bacci, S., and Gnaldi, M. (2016). *Statistical Analysis of Questionnaire: A Unified Approach Based on R and Stata*. Boca Raton, FL: Chapman & Hall/CRC Press.

Beaujean, A.A. (2018). Simulating data for clinical research: a tutorial. *Journal of Psychoeducational Assessment* 36: 7–20.

Bland, J.M. and Altman, D.G. (2011). Correlation in restricted ranges of data. *British Medical Journal* 342: d556.

Bradburn, N., Sudman, S., and Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires*, 2e. San Francisco, CA: Jossey-Bass Inc.

Brown, T.A. (2015). *Confirmatory Factor Analysis*, 2e. New York: Guilford Press.

Calvert, M., Blazeby, J., Altman, D.G. et al. (2013). Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 309: 814–822.

Calvert, M., Kyte, D., Mercieca-Bebber, R. et al. (2018). Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO extension. *JAMA* 319: 483–494.

Calvert, M., King, M.T., Mercieca-Bebber, R. et al. (2021). SPIRIT-PRO Extension explanation and elaboration: guidelines for inclusion of patient-reported outcomes in protocols of clinical trials. *BMJ Open* 11: e045105. https://doi.org/10.1136/bmjopen-2020-045105.

Cappelleri, J.C. and Bushmakin, A.G. (2014). Interpretation of patient-reported outcomes. *Statistical Methods in Medical Research* 23: 460–483.

Cappelleri, J.C. and Spielberg, S.P. (2015). Advances in clinical outcome assessments. *Therapeutic Innovation & Regulatory Science* 49: 780–782.

Cappelleri, J.C., Zou, K.H., Bushmakin, A.G. et al. (2013). *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*. Boca Raton, FL: Chapman & Hall/CRC Press.

Cappelleri, J.C., Lundy, J.J., and Hays, R.D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics* 36: 648–662.

Cella, D., Hahn, E.A., Jensen, S.E. et al. (2015). *Patient-Reproted Outcomes in Performance Measurement*. Research Triangle Park, NC: RTI Press.

Chassany, O., Sagnier, P., Marquis, P. et al. (2002). Patient-reported outcomes: The example of health-related quality of life – a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. *Drug Information Journal* 36: 209–238.

Chen, W.C., McLeod, L.D., Nelson, L.M. et al. (2014). Quantitative challenges facing patient-centered outcomes research. *Expert Review of Pharmacoeconomics & Outcomes Research* 14: 379–386.

Cook, K.F., Victorson, D.E., Cella, D. et al. (2015). Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. *Quality of Life Research* 24: 575–589.

Coon, C.D. and Cappelleri, J.C. (2016). Interpreting change in scores on patient-reported outcome instruments. *Therapeutic Innovation & Regulatory Science* 50: 22–29.

Coon, C.D. and Cook, K.F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research* 27: 33–40.

Coons, S.J., Gwaltney, C.J., Hays, R.D. et al. (2009). Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force Report. *Value in Health* 12: 419–429.

Creswell, J.W. and Clark, V.L.J. (2007). *Design and Conducting Mixed Methods Research*. Thousand Oaks, CA: SAGE Publications, Inc.

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadworth.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.

Crosby, R.D., Kolotkin, R.L., and Williams, G.R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology* 56: 295–407.

Crosby, R.D., Kolotkin, R.L., and Williams, G.R. (2004). An integrated methods to determine meaningful changes in health-related quality of life. *Journal of Clinical Epidemiology* 57: 153–1160.

de Vet, H.C.W., Terwee, C.B., Mokkink, L.B., and Knol, D.L. (2011). *Measurement in Medicine: A Practical Guide*. New York: Cambridge University Press.

DeMuro, C.D., Lewis, S.A., DiBenedetti, D.B. et al. (2012). Successful implementation of cognitive interviews in special populations. *Expert Review of Pharmacoeconomics & Outcomes Research* 12: 181–187.

DeVellis, R.F. (2017). *Scale Development: Theory and Applications*, 4e. Thousand Oaks, CA: SAGE Publications, Inc.

Edelen, M.O. and Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research* 16: 5–18.

EMA (European Medicines Agency), Committee for Medicinal Products for Human Use (2005). Reflection paper on the regulatory guidance for us of health-related quality of life (HRQOL) measures in the evaluation of medicinal products. European Medicines Agency. www.emea.europa.eu/pdfs/human/ewp/13939104en.pdf (accessed 15 April 2020).

EMA (European Medicines Agency), Committee for Medicinal Products for Human Use (2009). Qualification of novel methodologies for drug development: guidance to applicants. European Medicines Agency. www.ema.europa.eu/pdfs/human/biomarkers/7289408en.pdf (accessed 15 April 2020).

Fayers, P.M. and Machin, D. (2016). *Quality of Life: The Assessment, Analysis and Reporting of Patient-Reported Outcomes*, 3e. Chichester, UK: John Wiley & Sons Ltd.

FDA (Food and Drug Administration) (2009). Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. *Federal Register* 74 (235): 65132–65133. https://www.fda.gov/media/77832/download (accessed 15 April 2020).

FDA (Food and Drug Administration) (2019a). Clinical outcome assessments (COA): frequently asked questions. https://www.fda.gov/about-fda/clinical-outcome-assessments-coa-frequently-asked-questions (accessed 15 April 2020).

FDA (Food and Drug Administration) (2019b). Patient-focused drug development guidance series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making. Draft guidance documents. https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical (accessed 15 April 2020).

FDA (Food and Drug Administration) (2019c). Clinical outcome assessment (COA) qualification program. https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-coa-qualification-program (accessed 15 April 2020).

FDA (Food and Drug Administration) (2019d). Clinical outcome assessment (COA) compendium. https://www.fda.gov/media/130138/download (accessed 15 April 2020).

Feinberg, R.A. and Rubright, J.D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice* 35: 36–49.

Finch, H.W. (2020). *Exploratory Factor Analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

Fowler, F.J. Jr. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: SAGE Publications, Inc.

Giesinger, J.M., Loth, F.L.C., Aaronson, N.K. et al. (2020). Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research. *Journal of Clinical Epidemiology* 118: 1–8.

Glasser, B.G. and Strauss, A.L. (1999). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine Publishing Company.

Gnanasakthy, A., Mordin, M., Clark, M. et al. (2012). A review of patient-reported outcome labels in the United States: 2006–2010. *Value in Health* 15: 437–442.

Gnanasakthy, A., Mordin, M., Evans, E. et al. (2017). A review of patient-reported outcome labeling in the United States (2011–2015). *Value in Health* 20: 420–429.

Gnanasakthy, A., Barrett, A., Evans, E. et al. (2019). A review of patient-reported outcomes labeling for oncology drugs approved by the FDA and the EMA (2012–2016). *Value in Health* 22: 203–209.

Good, P.I. (2005). *Resampling Method: A Practical Guide to Data Analysis*, 3e. Boston, MA: Birkhäuser.

Hays, R.D. and Revicki, D. (2005). Reliability and validity (including responsiveness). In: *Assessing Quality of Life in Clinical Trials: Methods and Practice* (ed. P.M. Fayers and R.D. Hays), 25–39. Oxford, UK: Oxford University Press.

Johnson, R.B. and Christensen, L. (2017). *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. Thousand Oaks, CA: SAGE Publications, Inc.

Kerr, C., Nixon, A., and Wild, D. (2010). Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. *Expert Review of Pharmacoeconomics & Outcomes Research* 10: 269–281.

King, M.T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research* 11: 171–184.

Lasch, K.E., Marquis, P., Vigneux, M. et al. (2010). PRO development: rigorous qualitative research as the crucial foundation. *Quality of Life Research* 19: 1087–1096.

Mair, P. (2018). *Modern Psychometrics with R*. Switzerland: Springer International Publishing AG.

Marquis, P., Chassany, O., and Abetz, L. (2004). A comprehensive strategy for the interpretation of quality-of-life data based on existing methods. *Value in Health* 7: 93–104.

McGraw, K.O. and Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1: 30–46.

McHorney, C.A. and Tarlov, A.R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research* 4: 293–307.

McLeod, L.D., Fehnel, S.E., and Cappelleri, J.C. (2018). Patient-reported outcome measures: development and psychometric validation. In: *Biopharmaceutical Applied Statistics Symposium: Design of Clinical Trials*, vol. *1* (ed. K.E. Peace, D.-G. Chen and S. Menon), 317–346. Springer Nature Singapore Pvt Ltd: Singapore.

Mokkink, L.B., Terwee, C.B., Patrick, D.L. et al. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-related outcomes. *Journal of Clinical Epidemiology* 63: 737–745.

Mokkink, L.B., Terwee, C.B., Knol, D.L. et al. (2010b). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Medical Research Methodology* 10: 22.

Morris, T.P., White, I.R., and Crowther, M.J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 38: 2074–2102.

Nunnally, J.C. and Bernstein, I.H. (1994). *Psychometric Theory*, 3e. New York: McGraw-Hill.

Patrick, D.L., Burke, L.B., Gwaltney, C.J. et al. (2011a). Content validity – establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1 – eliciting concepts for a new PRO instrument. *Value in Health* 14: 967–977.

Patrick, D.L., Burke, L.B., Gwaltney, C.J. et al. (2011b). Content validity – establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 2 – assessing respondent understanding. *Value in Health* 14: 978–988.

Powers, J.H. III, Patrick, D.L., Walton, M.K. et al. (2017). Clinician-reported outcome assessment of treatment benefit: report of the ISPOR clinical outcome assessment emerging good practices task force. *Value in Health* 20: 2–14.

Presser, S., Rothgeb, J.M., Couper, M.P. et al. (2004). *Methods for Testing and Evaluating Survey Questions*. Hoboken, NJ: Wiley.

Prinsen, C.A.C., Vohra, S., Rose, M.R. et al. (2016). How to select outcome measurement instruments for outcomes included in "Core Outcome Set" – a practical guideline. *Trials* 17: 449.

Qin, S., Nelson, L., McLeod, L. et al. (2019). Assessing test-retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research* 28: 1029–1033.

Reeve, B.B., Wyrwich, K.W., Wu, A.W. et al. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research* 22: 1889–1905.

Revicki, D.A., Gnanasakthy, A., and Weinfurt, K. (2007a). Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: the PRO evidence dossier. *Quality of Life Research* 16: 717–723.

Revicki, D.A., Erickson, P.A., Sloan, J.A. et al. (2007b). Interpreting and reporting results based on patient-reported outcomes. *Value in Health* 10: S116–S124.

Revicki, D., Hays, R.D., Cella, D., and Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology* 61: 102–109.

Schuck, P. (2004). Assessing reproducibility for interval data in health-related quality of life questionnaires: which coefficients should be used? *Quality of Life Research* 13: 571–586.

Schwarz, N. and Sudman, S. (ed.) (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco, CA: Jossey-Bass Inc.

Stone, A.A., Turkkan, J.S., Bachrach, C.A. et al. (ed.) (2000). *The Science of Self-Report: Implications for Research and Practice*. Mahwah, NJ: Lawrence Erlbaum Associations.

Streiner, D.L., Norman, G.R., and Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to Their Development and Use*, 5e. New York: Oxford Univeristy Press.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass Publishers.

Teddlie, C. and Tashakkori, A. (2009). *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Thousand Oaks, CA: SAGE Publications, Inc.

Terwee, C.B., Bot, S.D.M., de Boer, M.R. et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology* 60: 34–42.

Terwee, C.B., Prinsen, C.A.C., Chiarotto, A. et al. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research* 27: 1147–1157.

Walters, S.J. (2009). *Quality of Life Outcomes in Clinical Trialsand Health-Care Evaluation: A Practical Guide to Analysis and Interpretation*. Chichester, UK: Wiley.

Walton, M.K., Powers, J.H. III, Hobart, J. et al. (2015). Clinical outcome assessments: conceptual foundation – report of the ISPOR clinical outcomes assessment – emerging good practices for outcomes research task force. *Value in Health* 18: 741–752.

Wicklin, R. (2013). *Simulating Data with SAS®*. Cary, NC: SAS Institute Inc.

Wild, D., Grove, A., Martin, M. et al. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation. *Value in Health* 8: 94–104.

Willis, G.B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: SAGE Publishing, Inc.

Willis, G.B. (2015). *Analysis of the Cognitive Interview in Questionnaire Design: Understanding Qualitative Research*. New York: Oxford University Press.