VALUE IN HEALTH 14 (2011) 978–988

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

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

Donald L. Patrick, PhD, MSPH[1,], Laurie B. Burke, RPh, MPH[2], Chad J. Gwaltney, PhD[3], Nancy Kline Leidy, PhD[4], Mona L. Martin, RN, MPA[5], Elizabeth Molsen[6], Lena Ring, PhD[7]*

[1]*Department of Health Services, University of Washington, Seattle, WA, USA;* [2]*Office of New Drugs, Center for Drug Evaluation Research, Food and Drug Administration, Silver Spring, MD, USA;* [3]*Department of Community Health, Brown University, Providence, RI, USA, and PRO Consulting, Pittsburgh, PA, USA;* [4]*United BioSource Corporation, Bethesda, MD, USA;* [5]*Health Research Associates, Inc., Seattle, WA, USA;* [6]*International Society for Pharmacoeconomics and Outcomes Research, Lawrenceville, NJ, USA;* [7]*Health Economics & Outcomes Research Division, AstraZeneca, Södertälje, Sweden, and Pharmaceutical Outcomes Research, Department of Pharmacy, Uppsala University, Uppsala, Sweden*

A B S T R A C T

The importance of content validity in developing patient reported outcomes (PRO) instruments is stressed by both the US Food and Drug Administration and the European Medicines Agency. Content validity is the extent to which an instrument measures the important aspects of concepts developers or users purport it to assess. A PRO instrument measures the concepts most relevant and important to a patient's condition and its treatment. For PRO instruments, items and domains as reflected in the scores of an instrument should be important to the target population and comprehensive with respect to patient concerns. Documentation of target population input in item generation, as well as evaluation of patient understanding through cognitive interviewing, can provide the evidence for content validity. Part 1 of this task force report covers elicitation of key concepts using qualitative focus groups and/or interviews to inform content and structure of a new PRO instrument. Building on qualitative interviews and focus groups used to elicit concepts, cognitive interviews help developers craft items that can be understood by respondents in the target population and can ultimately confirm that the final instrument is appropriate, comprehensive, and understandable in the target population. Part 2 details: 1) the methods for conducting cognitive interviews that address patient understanding of items, instructions, and response options; and 2) the methods for tracking item development through the various stages of research and preparing this tracking for submission to regulatory agencies. The task force report's two parts are meant to be read together. They are intended to offer suggestions for good practice in planning, executing, and documenting qualitative studies that are used to support the content validity of PRO instruments to be used in medical product evaluation.

Keywords: content validity, instrument development, patient-reported outcomes, qualitative research, regulatory.

Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

Background to the Task Force

During March 2009 the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Board of Directors approved the formation of the Patient Reported Outcomes (PRO) Content Validity Good Research Practices Task Force to develop a good research practices report to address methods for ensuring and documenting the content validity of newly developed PRO instruments to support medical product indications and labeling claims. This task force report extends the work of a previously published ISPOR PRO task force report on the use of existing or modified PRO instruments [1] that did not address how to establish and document content validity; that is, the specific methodologic practices involved in designing studies to gather evidence of content validity and the methods for evaluating and documenting content validity.

Researchers experienced in psychometrics and PRO instrument development working in academia, government, research organizations, and industry from North America and Europe were invited to join the task force leadership group. The task force met bimonthly to develop the topics to address, outline, and prepare the first draft report. Due to the large volume of information, the task force report was split into two parts. Part 1 [2] covers elicitation of key concepts using qualitative focus groups and/or interviews to inform content and structure of a new PRO instrument. Part 2 covers the instrument development process, the assessment of patient understanding of the draft instrument using cognitive interviews, and steps for instrument revision.

The task force authors presented their work to date at the ISPOR 15th Annual International Meeting during May 2010 in Orlando, Florida. In July 2010 the draft reports (Part 1 and Part 2), were

Authors listed in alphabetical order after lead author.

* *Address correspondence to:* Donald L. Patrick, University of Washington, Health Services, PO Box 359455 SeaQoL Group, Seattle, WA 98195-9455, USA.

E-mail: donald@u.washington.edu.

1098-3015/$36.00 – see front matter Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

doi:10.1016/j.jval.2011.06.013

sent for review to the nearly 400 ISPOR PRO Review Group members. The task force received many comments that were considered and addressed as appropriate. The task force authors presented their revised draft report for final verbal comments at the ISPOR 16th Annual International Meeting in Baltimore, Maryland, during May 2011. The revised draft report was sent for a final review to all ISPOR members during June 2011.

Collectively, the task force received 41 written reviews by 52 ISPOR members submitted individually or representing an organization. All written comments are published at the ISPOR Web site. A list of those members who commented is also available. For these comments, please go to the Evaluating and Documenting Content Validity for PRO Instruments link at the ISPOR Good Outcomes Research Practices index under the Patient Reported Outcomes heading at: http://www.ispor.org/workpaper/practices_index.asp or via the purple Research Tools menu at the top of the ISPOR homepage (www.ispor.org). All comments, many of which were substantive and constructive, were considered. Once consensus was reached by all authors on both drafts, the final report was submitted to *Value in Health* in July 2011.

Introduction

This second part of the task force report, assessing respondent understanding, builds on the instrument development practices described in Part 1 [2]. It describes the second phase of establishing and reporting evidence of content validity for a new PRO instrument by addressing the process of instrument and item crafting from the data gathered during concept elicitation interviews and the methods for gathering evidence that persons in the target population understand the instrument's structure and content. The article then reviews the design and conduct of cognitive interviews, discusses the decision-making process during instrument revision, and addresses documentation as it relates to content validity. It is important to note that items in a newly developed PRO instrument may be original and/or derived from existing instruments.

Cognitive interviewing follows the concept elicitation and item generation phase of instrument development and addresses two important issues. Based on respondents' answers during cognitive interviewing, it can be ensured that: 1) the instrument content captures the most important aspects of the concept(s) of interest; and 2) respondents understand how to complete the instrument, how to reference the correct recall period, the meaning of the items, how to use the response scales, and any other instrument features that may influence patient responses in the intended mode of administration.

Cognitive interviewing applies to all items considered for inclusion in the instrument that will be evaluated for its quantitative measurement properties and used to support claims in the clinical trials of the medical product. Evaluating the relevance of all items to respondents and the importance of items to concept measurement requires careful analysis of all available information, including patient input based on the qualitative data analyzed, developer experience and judgment, and input from content or therapeutic experts. Structured individual cognitive interviews are recommended with a broad range of respondents from the target population using techniques such as think aloud and/or verbal probing to ascertain exactly how an item is interpreted and a response formed. Results and conclusions can be influenced by respondent characteristics such as literacy, experience with the disease condition, or experience in completing questionnaires. Addressing all concerns and suggestions raised by every respondent is not possible and may not even be appropriate. Demonstration that the new PRO instrument is understandable to potential respondents, however, is an essential piece of evidence used by regulators for evaluating content validity.

Items should assess the entire continuum of severity, difficulty, or other response categories/levels relevant to the target population for clinical trials. Decisions to revise an item should be documented in an item-tracking matrix that also reflects the nature of the problems encountered and the decisions made in all rounds of cognitive interviewing.

Cognitive interviewing can minimize errors arising from respondent misunderstanding during data collection by assessing clarity of terminology, phrasing, and format. Cognitive interviewing methods allow testing of respondent understanding of the tasks required to complete a PRO questionnaire administered by any mode. Evidence should be obtained across modes of administration and language versions to demonstrate that respondents understand the instructions, items, and response options. Furthermore and most importantly, cognitive interviews can address the assumption in survey research that responses to questionnaire items represent a common understanding of item content and intent across respondents, permitting pooling of data for quantitative analysis in subsequent psychometric testing and clinical trials.

Two primary components are tested when evaluating if an instrument assesses the concept of interest and the patients' comprehension of the items in the questionnaire. First, the intent of the question: What do respondents believe the question is asking? Is this perception consistent with the intended meaning? The second component is the meaning of specific terms in the instrument. What do specific words and phrases in the instructions, items, and/or response options mean to respondents? Is that meaning consistent with the intent? Is it relevant to the concept of interest? Does it raise new content important to the concept of interest and/or new content not reflected in the instrument as designed? [3,4].

Figure 1 outlines five good practices related to the development of the instrument and the use of cognitive interviews for evaluating a new PRO instrument. Item crafting and/or item selection from existing instruments is based on good principles of item construction [5], with instructions and recall period also grounded in data from the concept elicitation interviews. A large number of

1. Develop items based on findings from concept elicitation
   - Develop criteria for item selection according to purpose of instrument and concept and conceptual framework
   - Select recall period and modes of administration
   - Draft instructions
   - Determine wording of each new question
   - Match each new item to response scale
   - Review items against item criteria
   - Select items for cognitive interviews
   - Determine readability
   - Determine order and sequence
   - Format the actual instrument for cognitive interviewing
2. Design cognitive interview process for the planned context of use
   - Identify population
   - Design cognitive interview process
   - Develop protocol and cognitive interview guide
3. Conduct cognitive interviews
   - Train interviewers
   - Train subject to think aloud
   - Use verbal probes
   - Monitor interview quality
   - Record and transcribe
   - Prepare result summaries
4. Make decisions to revise the patient-reported outcome instrument
   - Employ an iterative process
   - Reduce ambiguity in item language
   - Assess saturation
   - Balance respondent input with principles of item construction and decisions on conceptual framework
5. Document cognitive interview results for evaluation of content validity
   - Complete Item tracking matrix including final item, final response scale, any preliminary domain assignment, description of intent of item, and patient quotes supporting item intent

Fig. 1 – Five good practices in using cognitive interviews to evaluate patient understanding of a new patient-report outcome instrument.

concerns are considered in the item writing process to develop a preliminary draft of a new PRO instrument.

Following item construction, developers design and conduct cognitive interviews to evaluate and revise the instrument before piloting. In drug development, it is most efficient to complete cognitive interviewing before Phase II trials begin so that quantitative testing of the instrument and any changes in the measure can occur before confirmatory Phase III studies begin. The final good practice is summarizing the documentation of content validity evidence combining the results of all stages of qualitative research.

Good Practice 1: Create the Draft Instrument Based on Findings from Concept Elicitation

After concept elicitation and before cognitive interviewing, the PRO instrument is drafted with the context of measurement as well as the targeted claim in mind. The goal is to create a new measure with content, structure, and scoring that reflects the target concept and intended use; that is, to support clinical trial measurement objectives pertaining to treatment efficacy or safety in the target population. Content validity depends on the final instrument score reflecting the targeted concept.

The instrument development procedure involves multiple sources of information. It is an iterative process of drafting, evaluation, and revision. For example, to assess the frequency and severity of disease symptoms, factors considered include the known characteristic features of the target population specific to disease and demographics, patient input based on the qualitative data from the elicitation interviews in the same population, the potential suitability of drafted items capturing aspects of these symptoms, insight from clinicians who see patients in this population, insight from measurement experts experienced with these symptoms on item distributions in relation to symptom severity, and previous knowledge and experience in instrument development within and across therapeutic areas.

A draft instrument with candidate items is developed and subjected to cognitive interviewing with representatives from the target population. Revisions are made in the instrument, further interviews are conducted, additional revisions are made, and the procedure continues until an instrument suitable for quantitative evaluation is derived. Detailed consideration of instrument development [6] is beyond the scope of these articles on content validity, but important aspects of the development process affecting instrument drafting and examination of content validity are addressed briefly below.

Establish item criteria

Systematic decisions are required regarding the attributes of the selected item content, the appropriate recall period, the mode of administration, as well as the language and formatting to be used in the items. As shown in Figure 2, establishing criteria that can be used to guide and evaluate the item development process is a useful first step. These criteria are only illustrative. Criteria may be distinctive or atypical across different measure development projects. For example, items may be needed to capture the different severity levels of symptoms such as mild wheezing in asthma to severe and persistent cough in chronic obstructive respiratory disease.

Select concepts for inclusion in PRO instrument

The selection of content to include in a PRO instrument is accomplished by comparing the patient interview data (gathered according to the principles of concept elicitation described in Part 1), to expert input and studies in published literature. It may also be useful to assess the generalizability of a concept across patients and cultures. Because the percentage of patients who report a concept increases, so should the probability of including it in the PRO instrument.

Criteria	Item meets criteria Yes/No
The item captures the concept that is intended.	
The item is relevant to all members of the target population.	
The item is worded in a manner consistent with the expressions used by patients.	
The item reflects different levels of magnitude, e.g., severity, frequency.	
The item represents a single concept, rather than a multidimensional concept.	
The item is not likely to be vulnerable to ceiling or floor effects within the target population, i.e., it will change with treatment.	
The content of the items is appropriate for the recall period.	
The content of the item is appropriate for the mode of administration.	
The response scale corresponds to the stem.	

Fig. 2 – Sample criteria for evaluating new items.

For example, symptoms reported by 90% of patients may be more closely related to the pathophysiology of the disease under study than symptoms reported by 2% of patients. However, no universal or generalizable cutoff exists that can be used to make decisions; the instrument developer must decide what criterion is appropriate for the specific measurement context. Distinguishing spontaneous patient responses from those that result from specific probes may also help in selecting important concepts for inclusion [7]. How well the sample represents the concerns of patients who will participate in clinical trials is an important consideration during the instrument development process to ensure that the items and structure represent the diversity of patient experience. Patients with different levels of severity—for example, ulcerative colitis patients with severe depression—may consider depression an important aspect of their disease experience. Whereas other members of this target disease population with milder affective effects may not share this importance rating.

Consider recall period and mode of administration

In drafting items, recall period is an important consideration as it may differ by item content, saliency, and frequency of occurrence [8]. The interplay between recall period and content also varies by purpose of assessment, period of observation, and frequency of assessment [9]. In general, it is advisable to select a recall interval that is as short as possible [10] while balancing recall bias and respondent burden.

Single assessments covering a short period of time may not capture important aspects of patient experience. For example, asking patients only once about their pain at the current time may not provide a representative picture of the full range of the symptom or treatment experience. Instead, it may be desirable to ask about their current status multiple times to capture a valid and reliable estimate of their experience.

The variability and frequency of the targeted concept are also considered when establishing a recall period. As a general rule, events and experiences that are highly variable or happen frequently, such as symptoms associated with a chronic disease (e.g., acute episodes of diarrhea or frequent or uncontrolled micturations), are best measured via multiple, frequent assessments with short recall, such as daily diaries or ecological momentary assessments. Less variable or less frequent events and experiences may be captured intermittently using longer recall intervals [9].

Relatively rare yet salient events, such as coughing up blood, having seizures, or the occurrence of a migraine headache, may be

best assessed with measures involving longer recall intervals or with event-based diaries. Even if these types of events are rare, such signs or symptoms are relevant in a clinical trial setting where the effects of an intervention may be measured over several months. No single recall period will fit all applications; thus, a variety of factors such as saliency, frequency of occurrence, and respondent burden should be considered to optimize data quality and completeness [8].

Finally, the choice of recall interval can influence the selection of the best mode of administration. Instruments with shorter recall intervals, such as daily assessments, may need to be completed outside of the clinic via a patient diary. When patient diaries are used, it is important to ensure that patients complete assessments according to protocol and not, for example, immediately before a clinic visit [11]. In this context, electronic diary methods may be preferable to paper diaries. Electronic diaries can time- and date-stamp entries, allowing investigators to ensure that assessments were completed at the appropriate times. In a trial directly comparing the two modes, patients were compliant with more than 90% of electronic diary assessments, but only 11% of paper diary assessments [11]. Therefore, an electronic mode of administration may be preferable when asking patients about their experience across a short interval in the natural environment.

Mode of administration may also be related to the content of the PRO instrument. If an instrument is designed to assess sensitive topics, such as drug use or sexual behaviors, computerized assessments may be preferred over interviewer-administered assessments. For example, patients' reports of human immunodeficiency virus-related risk behaviors and symptoms may be more accurate on a computerized assessment than in response to an interviewer's questions [12].

Create the draft instrument text including match with response scale

Drafting an instrument includes writing the instructions that will be provided to the patient, the text of the items, and the response scales. Formatting the instrument is also part of the development process. Instructions orient a patient to the nature of the assessment (what is being assessed) and the time interval addressed (e.g., since you woke up, during the past 2 weeks). Failing to set a patient's expectations clearly in the instructions can result in patient confusion and invalid data.

Drafting specific items for an instrument involves identifying the dimensions or attributes of the concepts to be assessed. For example, in developing a pain instrument, the researcher must decide whether to measure intensity/severity, frequency, and/or the duration in relation to the specific pain concept chosen. As with the determination of which concepts to assess, determining which dimension of a concept to measure involves multiple sources of information. This includes clinical expertise, empiric data, and the qualitative data from the elicitation interviews with patients. Each is essential to decision making and documentation of content validity.

As noted in the description of the item criteria, each item or set of items in a PRO instrument should address a single concept and dimension of that concept (e.g., severity of knee pain). Ideally, the language used in the item reflects as closely as possible the language used by patients in the qualitative interviews. For example, although dyspnea may be the medical concept of interest to clinical experts, patients rarely use this term in their daily language. Similar concerns arise with the concept of fatigue. Fatigue may have a common meaning among clinicians and instrument developers, whereas it could be understood differently by patients depending on health status, language, education level, culture, or other factors.

The most appropriate terminology is determined during concept elicitation and documented with qualitative data from focus groups or interviews. Continuing with the example above, rather than using the term "dyspnea," patients will likely reveal that they use terms or phrases such as trouble breathing, shortness of breath, or breathlessness to represent the concept of dyspnea. Although fatigue may be mentioned and even understood by patients, other terms such as tiredness, weakness, or even sleepiness or exhaustion may be more valid, depending on the context. Therefore, PRO instruments should utilize the terms used and understood by patients from the target population with the targeted concepts in mind and the context of measurement.

Selecting a response scale is a critically important part of the item design process. There are multiple types of scales a researcher can choose from, including categorical (e.g., yes/no); Likert-type, numeric rating scales; and visual analog scales. In general, continuous scales may be more sensitive to treatment effects than categorical scales because categorical scales generally have increased variability in responses. No universally accepted response scale will suit every PRO instrument or every mode of administration. For example, some questions and response scales are less suited for use on various ePRO devices, such as items with extensive text on a small screen personal digital assistant or use of a visual analog scale on an interactive voice response system.

The advantages and disadvantages of different response scale options have been extensively discussed elsewhere [4,13]. Figure 3 displays sample items and response scales for possible use in item creation. An instrument developer must decide which type of scale is most appropriate for the concept to be measured; for example, symptom, impact, or another concept.

One good practice for item writers is to read aloud the stem of the item that may include recall period or other qualifiers, the item content, and the candidate response scale as complete sentences, using response scales described in Figure 3. Such oral rehearsal often elucidates awkward, mismatched, or potentially confusing item content, language, or flow.

A good practice in item creation is the consideration of cultural and linguistic issues, especially if the instrument has not been developed cross-culturally using simultaneous development in more than one language. Linguistic concerns can be minimized by asking an expert in PRO instrument translations to review the instrument for translatability. This expert can suggest adjustments that can be made before the instrument is finalized to improve its ease of translation and appropriateness for other cultures [14]. Culturally based idioms, such as "sleeping soundly" or "feeling blue," may not make sense in other languages and may require adjustments to achieve cultural and linguistic equivalence [15,16].

Evaluate items against item criteria

After items have been developed, they are put into an appropriate order and organized for clarity and ease of administration; for example, by domain or response option classification. The order of the items is important for enhancing patient understanding of the instrument. For example, it is advisable to group items with similar response scales or items that assess similar concepts together so that patients are not forced to switch back and forth between concepts and response scales as they complete the items.

Finally, the instrument is formatted for use in cognitive interviews. Formatting includes clarity of presentation and ease of administration. Issues to consider in formatting a PRO include placement of instructions; presentation of response options; use of numbers, boxes, or circles for recording item responses; positioning of page breaks relative to content; use of instructions to continue to the next page; placement of instructions beyond the first page; use of skip patterns; and font size and type [17,18]. Consideration should also be given to potential changes in mode of administration [19]. If a change from pen-and-paper to electronic

Example 1: Numeric Rating Scale (Average severity of pain; Dworkin et al. [27])

Please rate your pain by circling the number that best describes your pain on average in the past 24 hours.

0 1 2 3 4 5 6 7 8 9 10

No Pain — Pain As Bad As You Could Imagine

Example 2: Visual Analog Scale (Peak severity of nausea)

Place a mark (/) on the scale to indicate how you have felt over the past 24 hours.

At its worst, how severe was your nausea over the past 24 hours?

Not at all |----------| Extremely

Example 3: Verbal Rating/Adjectival Scale (Severity of dyspnea)

How severe was your shortness of breath over the past 24 hours?

No shortness of breath
Mild
Moderate
Severe

Example 4: Likert-type scale (Patient global impression of change)

Since starting the study medication, my knee pain is:

Very much improved
Much improved
Minimally improved
Unchanged
Minimally worse
Much worse
Very much worse

Example 5: Categorical scale (Occurrence of waking, due to dyspnea)

Did you wake up last night, due to shortness of breath? YES NO

Example 6: Numeric scale (Frequency of headaches)

How many migraine headaches did you experience in the past week?

Enter number: _____ headaches

Example 7: Amount of time (Duration of headache)

Think about the migraine headache that you had today. How long did it last?

Enter amount of time: _____ hour(s) and _____ minute(s)

Note: Adapted from Dworkin et al. [27].

Fig. 3 – Examples of response scales.

administration is likely, formatting should be consistent with the electronic mode to ensure clarity and ease of use in the latter form.

Good Practice 2: Design Cognitive Interview Process to Document Content Validity for the Planned Context of Use

Cognitive interviews have two goals. The first is to assess the respondent's comprehension of the questionnaire and its items in relation to their intended meaning. This evaluation includes the identification of any format or wording problems or difficulties with the instructions, item stems, recall period, or response options. Second is to evaluate comprehensiveness of content by checking that no important items are absent that would have an influence on the evaluation of the targeted concept. Cognitive interviews are critical to the content validity of the instrument because it is the last opportunity in this stage of development to make adjustments in the measure before it is administered to a large number of subjects for quantitative testing to determine the psychometric measurement properties of the instrument. Those studies should use the final version of the instrument that has confirmed evidence of content validity to avoid the need for replication of the quantitative testing. Content validity testing is incomplete without confirmation in well-designed cognitive interviewing studies that document that the scores produced by the instrument adequately represent the intended concept(s) and the item content, including the response options, recall period, and format, is appropriate, comprehensive, and understandable to the target population in the context of measurement targeted.

Study sample

Sample characteristics and size will vary depending on the target population and concept. Similar to the elicitation phase, participants in the cognitive interviewing phase are drawn from the target population with characteristics similar to those of patients who will be participating in the clinical trials. As is the case for all

qualitative research, the intent is not to generalize the results per se, but to make certain the instrument undergoes evaluation by people from the target population who can provide data consistent with the types of responses likely to occur in the population. With this in mind, it is advisable to recruit participants who would be considered typical or generally representative of the target population, as well as a purposive sample of those who may have unique responses or perspectives (e.g., specific disease characteristics) and those within the target population likely to have difficulty interpreting or completing the instrument (e.g., visual, reading, lower education levels, or other language difficulties).

Sample size requirements are variable. Although Willis [7] has suggested that seven to 10 interviews are sufficient to confirm patient understandability of the item, the number of interviews needed is a function of the complexity of the instrument, the diversity of the population of interest [20], and the number of questionnaire iterations necessary to fully explore patient understanding of items. The greater the complexity and diversity of the concepts being measured, the more likely it is that the instrument will require a larger sample size and several rounds of revision to yield confidence in content comprehensiveness, relevance, and respondent comprehension.

Good Practice 3: Conduct Cognitive Interviews

The most widely used cognitive theory model underpinning cognitive interviewing was developed by Tourangeau in 1984 [4]. It is a four-stage process to explain how information is stored, retrieved, and organized by respondents to answer survey questions. Using cough as an example in relation to PRO instruments, the four actions identified in this model are: comprehension of each question (i.e., experience of cough); retrieval of relevant information from memory (i.e., frequency and severity of cough); judgment of the information needed (i.e., when cough occurred and how intense or bad the cough seemed in relation to other experiences); and formation of the response (i.e., deciding which response option to endorse) [4,21].

Respondents first must understand or interpret in a consistent manner what the item or question is asking, then find the relevant information in their memory. They make a judgment about the information available to them, adapt the information to fit it to the question or expected response format, and finally they give their response to the question.

With these considerations in mind, preparation can be made for the conduct of cognitive interviews. A semistructured interview guide is used to direct the cognitive interview process. These interviews entail a different skill set than required for concept elicitation interviews and focus groups. Cognitive interviews are conducted using a pattern of questioning that may seem straightforward, but requires careful attention to verbal and nonverbal respondent cues about their perception of the items' content. Moreover, new issues may arise that were not apparent during the concept elicitation phase, requiring the interviewer to probe the information spontaneously for clarity and relevance to the new measure's content or structure. Figure 4 provides examples of poorly worded and preferred wording or probing to assess patient understanding and content coverage of a PRO instrument.

Train interviewers

The value of each cognitive interview in contributing to the documented content validity of the instrument is highly dependent on the interviewer's knowledge, sensitivity, empathy, and understanding of the goal of the interview process. Successful interviewers are knowledgeable, well organized, focused, and able to mentally track multiple issues and to speak clearly in a friendly manner while staying sensitive to a patient and any difficulty (s)he is undergoing.

New interviewers can benefit from an apprenticeship in which they observe more experienced interviewers in the field. Mock interviews can be used to refine the interview guide and train interviewers in the process and flow of the interview. They can assist the new interviewer in listening for respondent cues indicating understanding or misunderstanding of the task or content of the instrument and in pursuing lines of questioning to gain insight into perspectives useful in instrument revision. Interviewer effects are well known in survey research, and thus rigorous training is needed to limit the possibility of bias [13].

Considerations in performing interviews

The cognitive interview process requires respondents to think in different ways from their normal patterns. One particularly difficult cognitive task is the think aloud part of the interview. There is a long tradition of using the think aloud approach [22]. In the cognitive interview process, patients are asked to verbalize their thoughts as they work through the questionnaire—to think aloud or verbally articulate how they make sense of questionnaire items. The interviewer guides each patient through the items, first using think aloud for the item in focus. This is followed by various probes to establish that the item is understood. Their responses tell the interviewer about the intended concept as well. This technique can be used successfully with a wide range of PRO formats by using simple questions such as: "Tell me what you think this item is asking you about?"

Other questions or techniques may help identify any problems with the PRO questionnaire. Such questions include impressions about the relevance of the PRO items, identifying any general difficulties that might be affecting responses, identifying aspects of the concept that are not covered, and impressions about how long, difficult, or complex the questionnaire seems to understand the burden invoked on the patient by the questionnaire.

A new cognitive interviewing round is needed if problems are identified that result in revisions to the questionnaire. The size of each round may be a practical matter and will depend on the complexity of the instrument and the nature or magnitude of intervening changes. Careful instrument development and item drafting (Good Practice 1) help to reduce problems during the cognitive interviewing stage.

In addition to assessing the actual content match between what a patient offers and what the meaning is for the specific concept addressed by each item, cognitive interview results can be evaluated by examining the amount of spontaneous, relevant, and detail-specific answers participants offer. The extent to which clarifications are needed during the interview process may be a function of the interviewer, the draft instrument, the participants, or all of these factors.

The cognitive interview sessions are audio recorded and transcribed verbatim. Results can be presented as summaries of essential findings, including key think aloud or verbal probing quotations for each questionnaire item or concept. Problems with comprehension are evaluated for indications that changes might be needed, and any missing content that was identified during the interview can be noted with rationale. Figure 5 shows a small section of an example cognitive summary report.

Electronic PRO instruments

Electronic administration of PRO instruments often involves unique features that should be considered when designing cognitive interviews [19]. For example, early stages of electronic PRO development may present items on a single page to mimic a single screen so the patient cannot see previous or subsequent items. Either paper mock-ups or actual screen shots are sufficient in early

Purpose	Poorly worded	Preferred wording or probing
Instructions: To understand respondent's interpretation of the task (s) to be performed.	Are the instructions clear? Yes/No Are the instructions easy to read and understand? Yes/No	Can you tell me In your own words, what this instruction is asking you to do? Can you describe any confusion or difficulty you had in understanding these instructions? Are there any words or phrases that you would change to improve the instructions?
Recall: To identify how patients retrieve information, remember situations or events.	Is this recall period too long? Too short? Just right?	What does (timeframe) mean to you? Describe your experiences with [concept] over the (timeframe). What period of time did you think about when you were completing the questionnaire?
Item stem: To understand the clarity of the question from the respondent's perspective.	Do you like this question? Is this question clear? Is this question easy to understand?	What does [item content] mean to you? Using your own words, how would you explain what this question means?
Response options: To understand how participants interpret the response options and make decisions around response choice.	What response did you choose? Is this the best response for you?	Please read each response choice and tell me what it means to you. In thinking about your experience with [item x], which response best describes your experience? What caused you to choose this response? Would you ever choose A? Why or why not? Can you describe an experience where you might choose D?
Content coverage: To determine if the content in the instrument is comprehensive/to assure that there are no missing concepts.	Is the instrument comprehensive? Do the questions cover all aspects of [the concept]?	What other experiences do you have with [the concept] that are not covered in this questionnaire?
Format: To identify respondent difficulties with the presentation of the questionnaire or diary.	Is the format okay? Do you have any suggestions for improving the format? Were the skip patterns clear?	Observe the respondent completing the questionnaire. Note facial expressions, indications of reading difficulty, flipping pages or screens back and forth. Listen for comments about difficulty reading or questions that indicate lack of clarity or ease of use. Observe how the respondent completed this portion of the questionnaire. Note if skip patterns were correctly followed. What suggestions do you have for changing the questionnaire so it is easier to complete?
Length: To determine if the length of time it takes to complete the questionnaire is reasonable (does not burden subject).	Is the questionnaire too long? Too short?	What did you think about the amount of time it took you to complete the questionnaire?

Fig. 4 – Examples of poorly worded and preferred wording or probing for cognitive interviews to assess patient understanding and content coverage of a patient-reported outcome instrument.

Cognitive interview summary				
Item presented in cognitive interviews	Subject responses to inquiry about what item means	Subject responses to inquiry about difficulty with item	Comments and discussion	Suggestion for changes to item (action to take)
Item #: Overall, how severe was your plaque-related itching over the past 24 h?	ID#: How much did I itch in the last 24 h? ID#: When I read it, I was thinking about how bad the itching really is. ID#: It's asking me how severe the itching is on psoriasis was. ID#: It's asking me over the past 24 h how bad my itching was. ID#: It's asking how itchy I felt and how severe, like if I couldn't stop myself. ID#: It's just trying to get me to rate the itching, the severity of the itching.	ID#: No difficulty. ID#: Well, the itching comes from the scales; it doesn't actually come from the spots. ID#: No difficulty. ID#: No difficulty. ID#: No difficulty. ID#: No difficulty.	Patients generally understood the question to be asking about the severity of their itching related to their psoriasis. One patient commented that the itching did not come from the scales (plaques). A change to "psoriasis-related" may be needed to make the item more understandable to patients because there are differences in what patients think plaques are.	Suggested change: Overall, how severe was your psoriasis-related itching during the past 24 h?

*For illustrative purposes only; cognitive summary reports may take various forms.

Fig. 5 – Example of a small section of a cognitive summary report.*

stages as a means to test patient understanding of the instrument and make revisions as needed before the electronic programming is finalized.

Further cognitive testing on the electronic device itself is suggested before it is used for data collection. This is helpful because it provides an opportunity to identify any additional subject or site-related training requirements that might be needed to administer the new PRO instrument successfully in clinical trials. The term "usability testing" refers to the process of assessing respondent ability to use the software and hardware appropriately [19,23]. Usability does not evaluate respondent understanding of content, but rather augments cognitive interviewing by assessing the ease or difficulty with which respondents use any software or hardware that deliver item content.

Good Practice 4: Make Decisions to Revise the Pro Instrument

The cognitive interview protocol is designed to allow the developer to make an initial assessment of patient comprehension of the items as they were drafted, assess difficulties, decide on revisions to improve the items, assess the revisions with patients, and possibly revise and assess again. Decisions to make revisions are not always straightforward. If an initial group of cognitive interviews shows that four of the five participants were confused about the meaning of an item, it is fairly clear that work is required to improve that item so the concept can be clearly understood. Decisions are less clear when only one or two of the five participants have difficulty with the instructions, item stem, or response option.

Lack of clarity, misinterpretation, and unintended ambiguity are primary reasons for instrument revision. For example, the term "bothered" in an item stem might be interpreted as physical intensity of a symptom rather than the intended meaning of emotionally annoying or burdensome. Obvious difficulty or consistent confusion across multiple respondents with item structure or language calls for recrafting the items to achieve clarity. Frequent mismatches discovered between the items and the patient experience of the condition calls for a reconsideration of the concepts being measured, as well as the particular attribute being addressed (e.g., severity or frequency). It is sensible at this point to re-evaluate the concept elicitation data or reassess the representativeness of the sample groups participating in both phases of the instrument development process.

As the cognitive assessments and revisions continue, the goal is to reach a point at which there is sufficient evidence of no remaining problems with patient comprehension of the draft items. It can then be said that saturation has been achieved for the cognitive evaluation of the new instrument. There is no set number for how many interviews are required for sufficient evidence. This is a decision the developer makes while considering the complexity of the instrument, the amount and nature of revisions already made, and the relative heterogeneity of the interview sample.

Good Practice 5: Document Cognitive Interview Results for Evaluation of Content Validity

Although specific patient language and discussion of problems and issues can be captured in a summary table, the overall results of the cognitive process can be presented in number of ways. Several investigators have found it useful to present the number of revisions needed according to Tourangeau's model of four components (comprehension, retrieval, judgment, and response)

Concept name	Pain	Sleep impact	Emotional impact
Item No.	Item No.	Item No.	Item No.
Concept definition	Pain related to (condition)	Disturbance to sleep quality caused by condition-related pain	Emotional difficulties caused by condition related pain
Original item	Since you woke up this morning, how severe was your pain?	How many times did you wake up in the night because of your pain?	How worried have you been because of your pain?
Original item response options	0–10 scale (0 = not severe at all, 10 = as severe as I can imagine)	Enter number:____	0–10 scale (0 = not worried at all, 10 = as worried as I can imagine)
Attribute to measure	Severity	Frequency	Magnitude (of worry)
Change from first group of cognitive interviews	Since you woke up this morning, how severe was your pain at its worst?	How many times did you wake up last night because of your pain?	No changes in first group of cognitive interviews.
Rationale for change	Patients were not sure if they should think about their overall or most intense experience.	Patients reported seeing "in the night" as general and could mean "any night" as opposed to specifically "last night."	
Examples of patient quotations	*"I had pain several times today, some I would rate low because it didn't bother me so much, but one pain was really bad..."* *"I'm not sure if I should think about all pain in the day and average it, or just pick one I remember best to answer about…"*	*"I was thinking in an average night, how many times do I usually wake up"….* *"Most nights I only wake up once or twice."*	
Change from second group of cognitive interviews	No changes in second group of cognitive interviews.	No changes in second group of cognitive interviews.	How worried have you been today because of your pain?
Rationale for change			Some patients reported not being clear on how much time to consider when answering this item.
Examples of patient quotations			*"I'm not sure if its talking about pain just now, today or all week"*
Final item	Since you woke up this morning, how severe was your pain at its worst?	How many times did you wake up last night because of your pain?	How worried have you been today because of your pain?
Final response option	0–10 scale (0 = not severe at all, 10 = as severe as I can imagine)	Enter number:____	0–10 scale (0 = not worried at all, 10 = as worried as I can imagine)
Domain	Pain severity	Sleep disturbance due to pain	Worry due to pain
Intent of item	To assess how patients perceive the "badness" of their pain	To assess number of times respondent woke up from sleep because of pain (not other reasons)	To assess the extent respondent is worried about their pain for any reason

*For illustrative purposes only; item-tracking matrices may take various forms.

Fig. 6 – Example item-tracking matrix.*

[4,24,25]. Others have developed overview tables to identify the initial text, the final text, and the stage of the process in which the revision occurred. The details and rationale are discussed in their report text. The overall summary can be brief. Further documentation of the rationale for revising PRO items can be recorded in an item tracking matrix and linked to the actual item in its original and revised forms (Fig. 6). Whatever method of presentation is selected, communication of the overview of key findings of the cognitive process and actions taken is needed.

In the final report of the development of a new instrument, the item tracking matrix should include information that starts with the development of the item and continues through the compete finalization of the item. The information it provides is typically broad, including the final item and response category, how this item fits within the final conceptual framework of the instrument, and how the item was saturated based on cycles of interviews and links to patient quotes that were used in creating the item. It also contains all of the discarded items developed using concept elicitation as well as items that were either altered or dropped as a result of the cognitive interviewing process.

The item-tracking matrix may also be useful for cross-cultural adaptation where translators need to understand what is intended by every item and for linking claims to concepts to items and patient quotes supporting the item as well. A matrix that includes the intent of the item may be helpful in refining the claim language for medical products, too, pending quantitative validation of the new instrument and results of pivotal trials.

Using quantitative methods in content validity testing

In addition to the qualitative work, quantitative evaluation of items, such as assessment of how well items address the entire continuum of patient experience of the concept is useful and desirable, regardless of if the concept is a symptom, behavior, or feeling. Rasch analysis or item-response theory methods can be used to evaluate item information curves and what part of the response continuum items address [26,27]. The use of quantitative data in the absence of prior knowledge, frameworks, and qualitative considerations can lead to a theoretical instrumentation producing scores with unknown meaning. Similarly, the use of qualitative data alone to substantiate an instrument may be rhetorically convincing, but scientifically incomplete.

Conclusions

Eliciting concepts, crafting the instrument, conducting cognitive interviews in the target population, revising the instrument, finalizing the conceptual framework of the instrument, and documenting the methods and results of this iterative process complete the qualitative phase of instrument development. Building on qualitative interviews and focus groups used to elicit concepts, cognitive interviews help developers craft items that can be understood by respondents in the target population and can ultimately, confirm that the final instrument is appropriate, comprehensive, and understandable in the target population.

Conducting interviews with persons representing important subgroups of the target population, including age groups, sex, ethnicity, language group, socioeconomic status, literacy, and severity of condition, is often not feasible. However, every effort should be made to conduct cognitive interviews in a diverse sample representing the target population to ensure that the score produced by an instrument reflects the targeted concept.

Developing a PRO instrument with adequate content validity, as presented in both parts of this task force report, is an iterative process. During the development process, researchers may find that the concept initially targeted for evaluation is inconsistent with patient experiences or perspectives, resulting in the need to revise the target concept. Qualitative studies may change the developer's understanding of how data can be gathered using the instrument or how the quantitative data can be described or interpreted in terms of the target concept and population.

Qualitative research is specific to the targeted measurement concept in the context of measurement. If used to support a medical product claim, the concept of the claim needs to be consistent with the concept measured and the clinical trial objectives. If any revision in the targeted claim is considered, additional qualitative research would be needed before the PRO instrument can be deemed adequate to measure the new concept. Likewise, a change in the targeted context of use reflected by the target patient population and disease/condition would create the need to re-evaluate the content validity. Rigorous qualitative research provides evidence of content validity demonstrating the critical link between the measurement concept and the score produced by the instrument in a specific context of measurement. Content validity is critical for determining how clinical trial results derived from these PRO instruments can be described as claims in medical product labeling and advertising. But it is also critical for anyone—patients, clinicians, or others—who wants to know the meaning of clinical outcome data.

Acknowledgements

The authors thank Anne Skalicky for preparing the references, Kathy Hobson for preparing the tables and figures, Carrie Lancos for editing assistance, and the many ISPOR reviewers who spent considerable time and effort providing us with valuable comments.

REFERENCES

[1] Rothman M, Burke L, Erickson P, et al. Use of existing patient-reported outcome (PRO) instruments and their modification: ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report. Value Health 2009;12:1075–83.
[2] Patrick DL, Burke L, Gwaltney CJ, et al. 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 I – eliciting concepts for a new PRO instrument. Value Health 2011;14:967–77.
[3] Payne S. The Art of Asking Questions. Princeton, NJ: Princeton University Press, 1951.
[4] Tourangeau R. Cognitive science and survey methods. In: Jabine T, et al., eds. Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines. Washington DC: National Academy Press, 1984.
[5] Brislin RW. The wording and translation of research instruments. In: Lonner WJ, Berry J, eds. Field Methods in Cross-Cultural Research. Beverly Hills: Sage, 1996.
[6] Streiner D, Norman GR, Health Measurement Scales: A Practical Guide to Their Development and Use. 3rd ed. Oxford: Oxford University Press, 2003.
[7] Willis G. Cognitive Interviewing. Thousand Oaks, CA: Sage Publications, Inc., 2005.
[8] Norquist J, Girman C, Fehnel S, et al. International Society for Quality of Life Research ~ 2010 conference abstracts. Qual Life Res 2010;19:1–144.
[9] Stone AA, Atienza S, Nebeling L. The Science of Real-Time Data Capture: Self-Reports in Health Research. New York: Oxford University Press, 2007.
[10] US Food & Drug Administration guidance for industry—patient-reported outcome measures: use in medical product development to support labeling claims. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. [Accessed December 29, 2010].
[11] Stone AA, Shiffman S, Schwartz JE, et al. Patient non-compliance with paper diaries. BMJ 2002;324:1193–4.
[12] Locke SE, Kowaloff HB, Hoff RG, et al. Computer-based interview for screening blood donors for risk of HIV transmission. JAMA 1992;268:1301–5.
[13] Dillman D. Mail and Internet Surveys: The Tailored Design Method. New York: John Wiley and Sons, 2000.

[14] Conway K, Patrick D. Translatability assessment. In: International Workshop on Comparative Survey Design & Implementation. Available from: http://www.csdiworkshop.org/pdf/3mc2008_proceedings/session_01/Conway.pdf. [Accessed September 15, 2010].
[15] Wild D, Grove A, Martin M, et al. 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 Health 2005;8:94–104.
[16] Wild D, Eremenco S, Mear I, et al. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR patient-reported outcomes translation and linguistic validation good research practices task force report. Value Health 2009;12:430–40.
[17] Goldman M, Ram SS, Yi QL, et al. The donor health assessment questionnaire: potential for format change and computer-assisted self-interviews to improve donor attention. Transfusion 2007;47:1595–600.
[18] Mallen CD, Dunn KM, Thomas E, et al. Thicker paper and larger font increased response and completeness in a postal survey. J Clin Epidemiol 2008;61:1296–300.
[19] Coons SJ, Gwaltney CJ, Hays RD, et al. 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 Health 2009;12:419–29.
[20] Leidy NK, Vernon M. Perspectives on patient-reported outcomes: content validity and qualitative research in a changing clinical trial environment. Pharmacoeconomics 2008;26:363–70.
[21] Qureshi H, Rowlands O. User satisfaction surveys and cognitive question testing in the public sector: the case of personal social services in England. Int J Soc Res Method 2004;7:273–87.
[22] Kucan L, Beck IL. Thinking Aloud and Reading Comprehension Research: Inquiry, Instruction, and Social Interaction. Rev Educ Res 1997;67:271–99.
[23] Presser S, Couper MP, Lessler JT, et al. Methods for testing and evaluating survey questions. Public Opin Q 2004;68:109–30.
[24] Watt T, Rasmussen AK, Groenvold M, et al. Improving a newly developed patient-reported outcome for thyroid patients using cognitive interviewing. Qual Life Res 2008;17:1009–17.
[25] Ploughman M, Austin M, Stefanelli M, et al. Applying cognitive debriefing to pre-test patient-reported outcomes in older people with multiple sclerosis. Qual Life Res 2010;19:483–7.
[26] Andrich D. Rasch models for measurement. Beverly Hills, CA: Sage Publications, 1988.
[27] Hambleton R, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press, 1991.