Volume 12 • Number 8 • 2009
VALUE IN HEALTH

Use of Existing Patient-Reported Outcome (PRO) Instruments and Their Modification: The ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report

Margaret Rothman, PhD,[1] Laurie Burke, RPh, MPH,[2] Pennifer Erickson, PhD,[3] Nancy Kline Leidy, PhD,[4] Donald L. Patrick, PhD, MSPH,[5] Charles D. Petrie, PhD[6]

[1]WW Patient-Reported Outcomes Center of Excellence, Johnson & Johnson Pharmaceutical Services, LLC, Raritan, NJ, USA; [2]Study Endpoints & Labeling, Office of New Drugs, CDER, FDA, Silver Spring, MD, USA; [3]On-line Guide to Quality-of-Life Assessment (OLGA), State College, PA, USA; [4]Scientific Affairs, United BioSource Corporation, Bethesda, MD, USA; [5]University of Washington, Seattle, WA, USA; [6]Pfizer, Inc, New London, CT, USA

[Correction added after online publication 09-Oct-2009: Case Studies header added]

ABSTRACT

Background: Patient-reported outcome (PRO) instruments are used to evaluate the effect of medical products on how patients feel or function. This article presents the results of an ISPOR task force convened to address good clinical research practices for the use of existing or modified PRO instruments to support medical product labeling claims. The focus of the article is on content validity, with specific reference to existing or modified PRO instruments, because of the importance of content validity in selecting or modifying an existing PRO instrument and the lack of consensus in the research community regarding best practices for establishing and documenting this measurement property.

Methods: Topics addressed in the article include: definition and general description of content validity; PRO concept identification as the important first step in establishing content validity; instrument identification and the initial review process; key issues in qualitative methodology; and potential threats to content validity, with three case examples used to illustrate types of threats and how they might be resolved. A table of steps used to identify and evaluate an existing PRO instrument is provided, and figures are used to illustrate the meaning of content validity in relationship to instrument development and evaluation.

Results & Recommendations: Four important threats to content validity are identified: unclear conceptual match between the PRO instrument and the intended claim, lack of direct patient input into PRO item content from the target population in which the claim is desired, no evidence that the most relevant and important item content is contained in the instrument, and lack of documentation to support modifications to the PRO instrument. In some cases, careful review of the threats to content validity in a specific application may be reduced through additional well documented qualitative studies that specifically address the issue of concern.

Conclusion: Published evidence of the content validity of a PRO instrument for an intended application is often limited. Such evidence is, however, important to evaluating the adequacy of a PRO instrument for the intended application. This article provides an overview of key issues involved in assessing and documenting content validity as it relates to using existing instruments in the drug approval process.

Keywords: content validity, instruments, outcomes, patient-reported outcomes, validity.

Background to the Task Force

In January 2007, the ISPOR Health Science Policy Council recommended to the ISPOR Board of Directors that an ISPOR Patient-Reported Outcomes (PRO) Task Force on Use of Existing Instruments and their Modification be established. The Board of Directors approved this PRO Task Force in March 2007. The PRO Task Force chair and members were chosen based on their experience as scientific leaders in the field, as well as developers and users of PRO instruments. A range of perspectives on PRO instruments was provided by the diversity of their work experience: research organizations, government, academia, and industry.

The PRO Task Force met every 6 weeks to develop the report outline and discuss issues that arose in the manuscript's development. The manuscript outline was presented for comment at ISPOR 12th Annual International Meeting, Arlington, VA, USA in 2007. The first draft report was presented for comment at the 13th Annual International Meeting, Toronto, Canada in 2008.

In March 2009, the second draft report was submitted for review and comments to the 360 members of the ISPOR PRO Special Interest Group. Comments were discussed and incorporated. A revised draft report was presented for final comment at the ISPOR 14th Annual International Meeting, Orlando, FL, USA in May 2009. The task force then addressed and incorporated comments elicited during this final review process as appropriate. All written comments on the PRO Existing Instruments and Their Modification Task Force Report are available on the ISPOR Web site at: http://www.ispor.org/TaskForces/PROInstrumentsUse.asp.

Address correspondence to: Margaret Rothman, PhD, WW PRO COE, JJPS, 301 South Alexander Avenue, Washington, GA 30673, USA. E-mail: mrothman@its.jnj.com

10.1111/j.1524-4733.2009.00603.x

Introduction

During early deliberations, the task force discussed the importance of content validity in selecting or developing a modified

Table 1 Steps in identifying and evaluating an existing PRO measure

☑	Step
☐	1. Name and define the concept.
☐	2. Draft the claim, target population, target product profile, and end point model.
☐	3. Identify candidate measures.
☐	4. Identify or formulate a conceptual framework for the instrument(s).
☐	5. Assemble and evaluate information on development methods.
☐	a. Elicitation focus groups and interviews; sample size and characteristics relative to intended use; analytical approach; results, including evidence of saturation
☐	b. Cognitive interviews; sample size and characteristics relative to intended use; methods (mode of administration); results
☐	c. Transcripts—for independent review and stratified analyses; evidence of saturation
☐	6. Conduct any needed qualitative research.
☐	a. Prepare documents for decision-making and regulatory submission—including protocol and study report, saturation tables, transcripts.
☐	b. Map qualitative data to existing instrument items.
☐	7. Assess adequacy of content validity for purpose.
☐	a. Identification of any concepts, domains, or items relevant to patients in the target population not included in the existing instrument
☐	b. Assess the relevance and importance of this content
☐	8. Determine the need for modifications or new instrument development.
☐	a. Evaluate cost benefit in terms of need, timelines, and resource allocation.
☐	b. Discuss alternatives—change the claim/concept; alter PRO positioning.

PRO, patient-reported outcome.

version of an existing PRO instrument and the lack of consensus in the research community regarding best practices for establishing and documenting this important measurement property. The task force decided to focus their review and discussion on content validity in PRO evaluation with specific reference to existing or modified instruments. It should be noted that this article is meant to represent current best practices in assessing the content validity of existing PRO instruments for purposes of making a regulatory claim rather than a repetition of the Food and Drug Administration (FDA) Draft PRO Guidance [1]. We feel that the recommendations in this article are consistent with the guidance; however, terminology and interpretation may vary slightly. A working knowledge of instrument development and qualitative methods is assumed; those readers who desire to know more about this topic are encouraged to read one of the many basic textbooks, including those referenced throughout the document. The article begins with a definition and general description of content validity. This is followed by a discussion of PRO concept identification as the important first step in establishing content validity in the context of a regulated claim (section III). A description of identification of an instrument and the initial review process, summarized in Table 1 (section IV), is followed by a discussion of key issues in qualitative methodology, the foundation of content validity (section V). The article concludes with a discussion of potential threats to content validity (section VI), with three case examples used to illustrate types of threats and how they might be resolved, as presented in Tables 2 and 3. References of existing work in each of the areas discussed are provided throughout the article to guide the reader interested in further information. Whereas an overview of key issues in qualitative research methodology as it relates to either selecting or developing a modified version of an existing instrument is presented, it is not the intent of this article to provide an in-depth review of these methods which is available in more detail elsewhere.

Content Validity: Basic Principles

Definition: The goal of measurement is to quantify a concept. *The Standards for Educational and Psychological Testing*, a key reference describing development and evaluation of instruments used to evaluate an individual's behavior, defines validity as an overall assessment of the degree to which evidence and theory support the interpretation of scores entailed by proposed uses of the instrument [1]. (*The Standards for Educational and Psychological Testing* uses the term "tests" rather than "instruments." The term "test" may be confusing when used in the context of PRO research; therefore, the term PRO "instruments" is used in this document.) One type of validity is based on content. Evidence of content validity is obtained from an analysis of the relationship between an instrument's content and the construct it intends to measure. *The Standards for Educational and Psychological Testing* uses the term "construct" as "the concept or characteristic that a test is designed to measure." The term "concept" will be used in this article to be consistent with the

Table 2 Framework (model) for evaluating content validity of an existing instrument within the context of a specific claim

Element*	Acceptable level of agreement†	Steps to remediate (examples)	Documentation of evidence (resources and examples)
1. Conceptual match‡			
2. Input from the target population§			
3. Item content, including saturation¶			
4. Modification, e.g., mode of administration, translation**			

*The greater the absence of evidence to substantiate content validity, the more comprehensive the evaluation and the greater the degree of remediation required. Definitions of the elements follow:
†Level of evidence refers to the basis on which the elements are evaluated. The following four levels of evidence are proposed: completely met, mostly met, partially met, not met.
‡Conceptual match: The concepts, as defined by the developer and represented by the items in the instrument, match the intended claim.
§Input from the target population: Patient concerns were obtained using appropriate qualitative methods.
¶Item content: Each concept in an item reflects patient concerns across the range of patients appropriate for the intended claim.
**Modification: Any change in the instrument from the original version needs to be identified and evaluated for its impact on content validity in terms of the above elements, in addition to the psychometric performance of the modified version relative to the original form.

Table 3 Case examples: threats to validity

Scenario A: use of an existing PRO instrument in a new patient population

A PRO measure designed to assess dyspnea is used routinely by pulmonologists in their clinical practice for evaluation and management of patients who have compromised lung capacity because of various etiologies. This standardized measure was developed and validated in a COPD patient sample.

Question: Does the development and use of this standardized measure in clinical practice support its use in a clinical trial that is being designed to assess experiences of shortness of breath in persons with asthma, with the ultimate goal of obtaining a label claim for dyspnea?

Issues/threats to content validity:

1. Is the concept of dyspnea, as expressed in the existing instrument, appropriate for persons with asthma?
2. Was the instrument developed based on patient input (e.g., focus groups)?
3. What are the implications of using an instrument developed for managing patient care in a clinical trial?

Approaches to remediation

1. Conduct qualitative analyses to gain understanding of how asthma patients conceptualize dyspnea.
2. Do an extensive literature to determine whether or not patient input has been included either in the instrument's development or subsequent application.
3. Compare and contrast clinical trial inclusion and exclusion criteria with characteristics of patients in a typical pulmonology practice.

Scenario B: short form of an existing PRO instrument

A multi-item, multiscale PRO measure has been reduced to a shorter, clinically efficient assessment form using accepted item response theory (IRT) psychometric procedures, including cross-validation in the relevant patient samples, and has been shown to have nearly identical psychometric properties, including factor structure, to the original long form of the instrument.

Question: Does such a shortened measure pose a threat to content validity for a regulated claim when used in the same patient population in which it was originally developed?

Issue/threat to content validity:

Did the process of generating the short form eliminate items that then, in turn, change at least one domain that measures a concept that may be essential to the specific language of the intended claim?

Approach to remediation

1. Conduct patient interviews or focus groups consisting of representative patients to determine the importance that patients placed on the omitted items relative to those retained in the shorter version.

Scenario C: focus group appraisal of PRO measure identifies same concepts with different item wording

The sponsor has selected an existing instrument that reportedly has the claim-relevant concept(s), and has been developed for the same population as the intended study group and decides to verify content validity in a small number of patient focus groups. From these groups, the sponsor learns that the item content is confirmed as relevant and that the instrument captures the range of experience of the construct being assessed. There does seem to be some differences, however, in the use by some patients to describe their experience.

Question: Are the different words used by a few of the focus group patients sufficiently synonymous in terms of a conceptual match to confirm the content validity of the original measure?

Issues/threats to content validity:

1. Do the different words used to describe the patient experience truly represent a different aspect of the concept, or is it essentially the same concept being described with different words?
2. What importance is placed on the number of focus group patients who use different wording (i.e., if one patient, is it cause for concern)? What about two patients, etc.?
3. Are the differences in wording in the instrument compared to that used by the focus group patients sufficient to require adjustment in the original measure?

Approaches to remediation

1. Conduct cognitive interviews to assure patient understanding consistent with concept. If the concept is consistent, no additional remediation is required.
2. If the interviews uncover patient misunderstanding, modification in the instrument would be required.

COPD, chronic obstructive pulmonary disease; PRO, patient-reported outcome.

FDA draft PRO guidance. Content refers to the themes or subjects addressed in the instrument; the wording and format of items, tasks, or questions on an instrument; as well as the guidelines for procedures for administration and scoring. In the context of content validity, the appropriateness of the content is related to the specific inferences to be made from the instrument scores. An evaluation of content validity is therefore essential in the selection of instruments to evaluate PROs to be used in making labeling claims.

The classic text, *Psychometric Theory*, by Nunnally and Bernstein [2], notes that there are two major standards for ensuring content validity. One standard is the representative nature of the collection of items comprising the instrument. Because "random sampling" is impractical, the method used to identify and select the items to represent the concept must be explicit. The second, related standard is based on the methods used in instrument construction, i.e., the rigor with which the instrument is constructed or formulated, including item and response option wording, scaling method (e.g., dichotomous, Likert, visual analog), and organization [1]. The appropriateness of a given content domain is related to the specific inferences to be made from the instrument scores [1].

PRO instruments are designed to capture data related to the health experiences of individuals, specifically the way patients feel or function in relationship to their condition or disease and/or their treatment. Figure 1 depicts the relationship between disease attributes, including observed signs and laboratory values, and patient-reported manifestations of the condition, e.g., symptoms, and patient experiences, including their descriptions of the disease and human experiences unrelated to the disease. For any disease, there are a host of characteristics, including observed signs, laboratory values, and patient-reported manifestations of the disease. Patients with the disease have a wide range of experiences, including those directly related to the disease itself, and other experiences that may be important to the patient's life experience, but are not characteristic of the specific disease of interest. For PRO instruments, content validity begins with the intersection between disease characteristics and patient experience, as shown in Figure 1, and is evaluated in terms of the adequacy with which this intersection "universe" of content is sampled in a given instrument to accurately capture patient-reported manifestations of the disease. Although the focus of this article is on PRO instruments used to support evidence of treatment efficacy, the adequacy of content validity for instruments to assess adverse treatment impact would follow the same principles.

In practice, content validity is determined by the relationship between the intended measurement concept and the methods used: 1) to develop and select items; 2) to evaluate the content; and 3) to formulate the instrument. A detailed description of these methods and their results provides evidence that the pro-

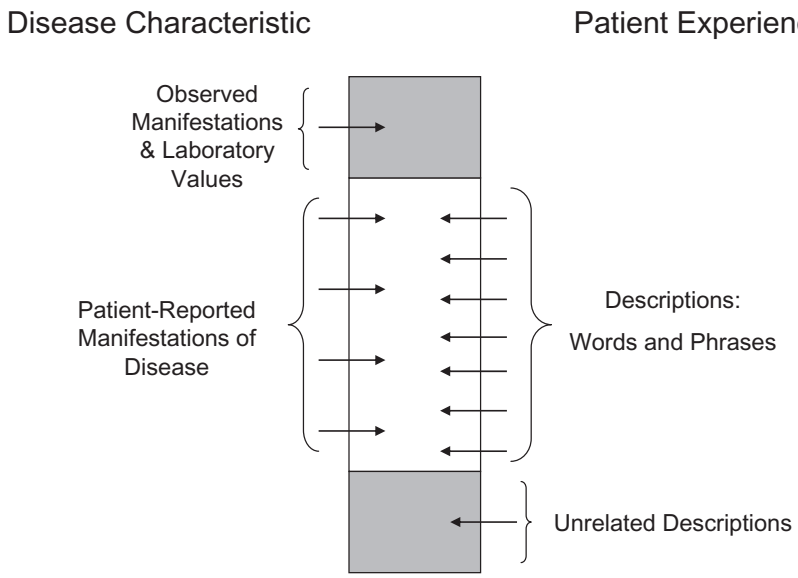

Figure 1 Content validity: the intersection of disease characteristics and patient experience.

posed use and interpretation of scores from the instrument represent the intended concept and therefore possess content validity [1–5]. Documentation of content validity describes the following three processes: identifying and defining the concept of interest, identifying the patient experience related to the targeted concept, and using appropriate methodology to develop the instrument to capture and quantify the concept.

Concept Identification within the Labeling Context

Selecting or developing a modified version of an existing PRO instrument for evaluating treatment efficacy begins with the identification of the concept to be measured and its positioning in relationship to other trial end points to be used in the development program, as described in the United States FDA PRO Draft Guidance [6]. Outcomes essential for testing product efficacy and informing appropriate use in clinical practice are considered highest priority and serve as primary end points in clinical trials. Outcomes considered important for providing additional understanding of treatment benefit or for communicating product attributes of value to the patient are also identified early in the development program and can serve as key secondary end points. Exploratory end points are useful for product development planning, including gathering data to inform future trial design decisions, but are not used to support labeling or promotional claims. PRO instruments can serve as primary, secondary, or exploratory end points.

Targeting the desired claim using the target product profile (TPP) approach can be useful when selecting and positioning the PRO, and linking it to the desired claim [7,8]. An end point model that describes each study end point, including PRO end points and their relationship to other end points, may be developed consistent with the goals identified in the TPP. An end point model should clearly delineate the relative importance of each end point in terms of labeling priorities to inform the subsequent drafting of study objectives, study protocol, and statistical analysis plan. Each concept to be measured is included, as well as how each concept will correspond to the ultimate labeling goals (i.e., the claims).

Selecting and defining the concept, specifying the intended claim, identifying the target population, and drafting the TPP are the initial steps in assuring and documenting content validity. Existing instruments are identified and reviewed with this information in mind.

Instrument Identification and Initial Review Process

Existing PRO instruments can be identified through literature searches, Web searches, and dedicated instrument databases. A review of candidate instruments includes a close examination of the items (stem and response options), recall period, mode of administration and instructions in relationship to the targeted concept in keeping with good scientific practice [9], and the Draft PRO Guidance [6].

If a conceptual framework of the instrument is available, it is examined for consistency with the concept. A conceptual framework is a detailed description or diagram of the relationship among the concepts, domains, and items comprising the instrument [10,11]. If such a framework does not exist, one should be developed, showing how the items, subscales, and total scales are related to one another and to the underlying concept and claim. The names used to describe the concept and subscales should be critically evaluated in light of the content and structure of the items and the targeted PRO claim. Adjustments in the name or concept referenced in the PRO instrument may be desirable to more accurately reflect the content and link to the claim. These adjustments, however, should be made after consulting with the instrument developer, keeping in mind that any name changes can have an adverse effect on subsequent interpretation and attempts to replicate findings. Strong and clear links between item content, subscale names, concept names, study objectives, and target claims are desirable and enable ease of understanding, interpretation, and communication.

A complete understanding of the methods used to develop a PRO instrument is essential to evaluate the suitability of an existing PRO instrument for any purpose. These methods are generally available in published literature and documentation; however, some may need to be obtained directly from the developer. An instrument may be best evaluated for potential use to support claims if: 1) patient-derived qualitative data forming the basis of the instrument are available; 2) a careful critique shows

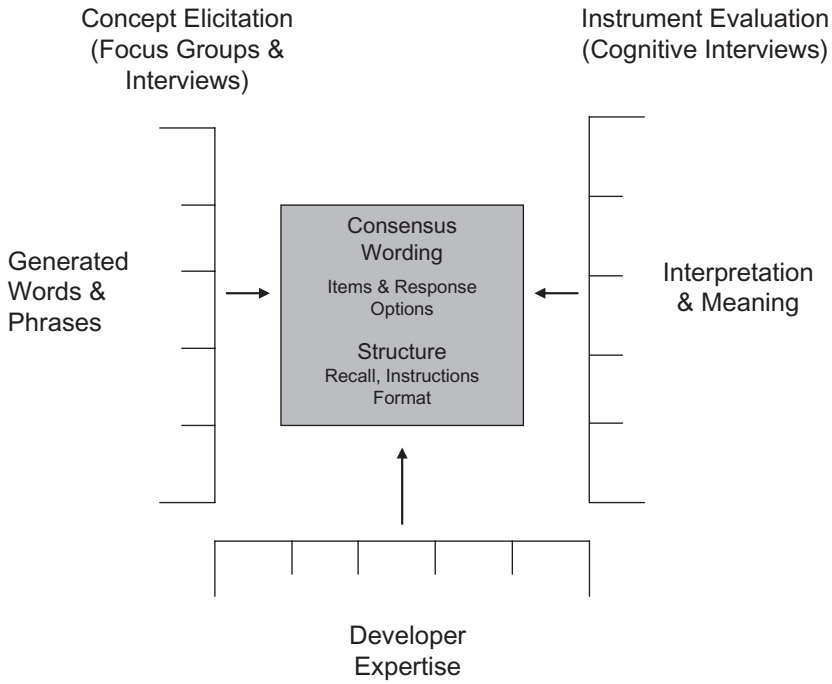

Figure 2 Content validity: content consensus through qualitative research.

the data generation methodology to be sound, and a sample similar to the development program's target population was used; and 3) the results of this qualitative work are comprehensive and relevant. Documentation of methods and results should demonstrate that the instrument is adequate for decision-making, and appropriate for regulatory submission as part of a PRO evidence dossier [12].

It is not uncommon for existing instruments to have little to no published or available information on their development history, to have little to no patient input in the development process, or to have minimal qualitative data from the patient population specifically enrolled in the trial. In these cases, there is little empirical evidence for the sponsor or regulatory reviewer to make an informed decision regarding the content validity of a PRO instrument, and suitability for the drug development program. If clinical and research experts conclude that the items in the PRO instrument may be an adequate representation of the concept in the intended claim and that there is quantitative evidence of reliability and validity, then qualitative research methods can be used to evaluate and substantiate the instrument's suitability as a trial outcome for labeling claim intent. If it is decided that changes to an instrument may be beneficial to the content validity for a particular purpose, it is recommended that the researcher discuss such changes with the developer as a professional courtesy, to gather new information as a potential colleague in the effort, and to avoid violations associated with copyright law. Because qualitative methods are essential to selecting and documenting the content validity of an existing instrument and to performing content valid modifications if necessary, the following section provides an overview of key aspects of the methods particularly relevant to evaluating existing PRO instruments to support regulated claims.

Qualitative Methods: The Foundation of Content Validity

"Qualitative research methods involve the systematic collection, organization and interpretation of textual materials from talk or observation. It is used in the exploration of meanings of social phenomena as experienced by individuals themselves, in their natural context" [13]. These empirical methods, specifically focus groups and 1:1 interviews, are used to elicit information from patients to inform instrument development. That is, these methods are used to obtain descriptive words and phrases, as shown in Figure 1. Cognitive debriefing interviews, generally conducted 1:1, are used to evaluate patient understanding of an instrument, as a draft item pool or as an existing measure being evaluated for possible use. Items are developed and evaluated with the target patient population in mind with wording designed to maximize ease of reading, translatability, and content coverage, with the final instrument reflecting all of these considerations. As shown in Figure 2, a PRO instrument is based on the concepts, words, and phrases generated during elicitation focus groups and interviews, with adjustments made based on interpretation and meaning provided during cognitive interviews with a new set of patients. Although patients are experts in their personal experience with a disease, they are not experts in the disease pathophysiology or instrument design techniques. Instrument developers play a key role in the development of consensus wording of items and response options, the selection of recall period, and the instructions to respondents, based on their knowledge of the content area and instrument development techniques.

Sample Selection

For product labeling in the context of a regulated claim, the FDA Draft PRO Guidance [6] indicates that PRO instruments assessing treatment benefit should show evidence of content validity using methods that include the elicitation of input from the patient population for which the claim is intended, and results that demonstrate the relevance and importance of item content to this group. Because these instruments are designed to capture patient experiences, the Draft PRO Guidance suggests this input be elicited directly from the patients targeted for the clinical trials. Clinicians or other experts or literature may be useful in

preparing interview guides and in defining enrollment criteria in the qualitative studies, but without patient input, PRO instrument development is incomplete.

Patients should represent variations in severity of the target condition, as well as in population characteristics such as age, sex, ethnicity, and language groups in accordance with the anticipated characteristics of the patients to be enrolled in the clinical trials. For example, the content of an existing instrument may initially have been generated on a broad population of adults ranging from age 18 to 65 years. Subsequent use of the instrument outside of this age range, e.g., adolescents or older adults, would require additional qualitative research for eliciting or confirming age-relevant content, testing modifications of the existing item content, and determining adequate comprehension of the instrument items and response format.

Data Collection Methods

Data from 1:1 interviews and/or focus groups with patients form the basis of PRO instrument content. When evaluating candidate instruments, users should examine the data collection methods used to generate the instrument to understand the content validity of the instrument. Gathering qualitative data through 1:1 interviews or focus groups is both a rigorous scientific method requiring a well-defined protocol, and an art requiring trained and experienced interviewers or focus group moderators. Interviews are audio recorded for later transcription and analyses. To assure representation from all group participants and to assist in data analyses, focus group interviewers/moderators are trained to encourage rapport and elicit comments from all participants. Focus groups can be videotaped and/or audio recorded. An assistant moderator can take notes with participant initials, and key words or quotations to facilitate data transcription and analyses. This individual can also map the discussion, marking the frequency with which various participants contribute comments to the discussion, and alerting the moderator for the need to query certain participants who have been less active in the discussion.

One-on-one interviews require a skilled interviewer and are particularly effective for sensitive topics unsuited to a group discussion, e.g., urinary incontinence, or for patient populations unable to participate or uncomfortable in a group setting, e.g., men with erectile dysfunction. One-on-one interviews are also used for cognitive interviews in which patients review an existing instrument or item pool, and can provide the developer with both insight into the extent to which their interpretations match the intent of the items and with any critical content that has been omitted from the instrument.

Cognitive debriefing interviews with patients from the target population are used to evaluate patient understanding of the items relative to the concept of interest. Also known as complement elicitation focus groups or interviews, cognitive interviews provide additional evidence of content validity. Cognitive interviews also provide an opportunity to query patients about the comprehensiveness of the instrument content relative to their experiences with the concept of interest, serving as a "pilot test," to make certain the instrument selected is, in fact, interpreted correctly, no additional instructions or special training is required, and that all of the appropriate concepts are covered. Specifically, at the end of the interview, patients may be asked if there were any aspects of the concept, e.g., experiences, symptoms, or sensations that were not addressed in the instrument, and if so, how important these are to the concept. If missing themes emerge across multiple interviews, and these themes are clearly related to the underlying concept, it is likely the instrument is missing important content and should be modified before use in a development program. This finding is referred to as "construct underrepresentation" [1].

Qualitative data from elicitation or evaluative methods not only provide information on the content validity of an existing measure, but also offer insight into concept names used to represent scales or subscales. As discussed previously, an assessment of the formulation of the conceptual framework of an instrument may suggest that scale or subscale names created by developers are unclear or inaccurate, particularly for regulated claims. Qualitative data can inform the evaluation of instrument naming conventions, and suggest alternatives more suited to the concept and proposed claim. Although it is usually inappropriate to rename existing instruments and domains, an accurate description of the content may facilitate communication between researcher and reviewer.

Data Analysis

Analyses of qualitative data use a carefully constructed methodology that includes independent coding with inter-rater analyses and reconciliation. When reviewing instrument development methods, attention should be paid to how the data were analyzed. Coding transcripts by participant, using initials or other coding system to protect anonymity, allows the researchers and reviewers to evaluate the representativeness of content across participants, and provides assurance of saturation (discussed in greater detail below). Coding by a specific patient characteristic, such as gender or disease severity, permits stratified analyses with an assessment of the consistency in experiences, and the words and phrases to describe these experiences by patient subgroups. Reference to this type of analysis can reassure sponsors and regulators that the instrument development methods were rigorous and that the instrument captures the content most relevant for the concept across the full range of the target population.

Analyses of focus group and interview data to evaluate and document the content validity of an existing measure are similar to those used in instrument development, identifying themes that emerge from the data in relationship to the concept of interest. These themes are used as analytical codes that are then mapped to the existing instrument content, with words and phrases compared with the wording used in the instrument. Figure 3 shows three examples of the relationship between the content of an existing measure relative to the universe of content derived through qualitative research methods. Example A shows a strong match between the content in the instrument and the information provided by patients. Note that the instrument content is not all inclusive, but represents the vast majority of the potential item content. Example B shows a poor match, with the content capturing less than 30% of the possible concept content. Example C shows a mismatch, with some of the instruments covering the content of interest, and coverage of content external to the concept of interest. It is unlikely that the instruments in examples B and C would be suitable for use as a primary or key secondary end point in registration trials because of the inadequate coverage of item content in the instrument with relevant information provided by patients. In such cases, a decision must be made either to adapt the existing instrument or to develop a new one, or to re-evaluate the intended claim and relevant PRO concept to support it.

Saturation

Qualitative data should be gathered to the point of saturation to ensure that the items in an instrument appropriately represent the relevant "universe of content" for the concept when conducting focus groups or 1:1 interviews. In instrument development, satu-

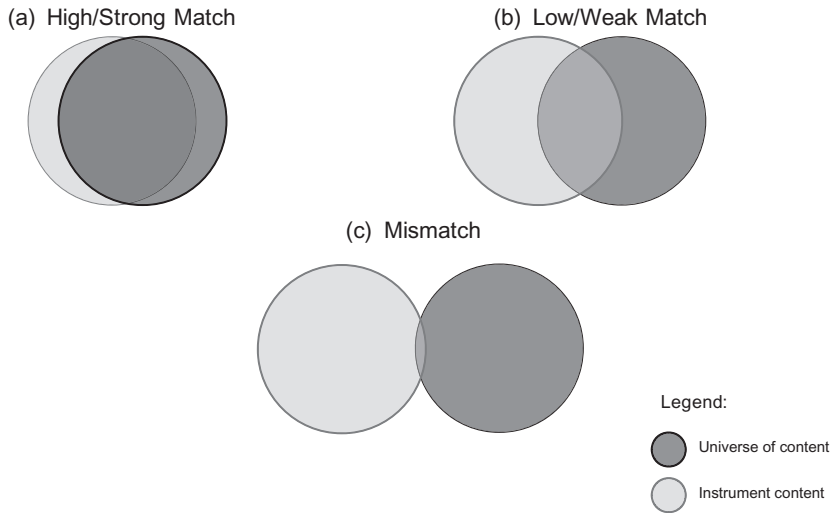

Figure 3 Instrument content versus universe of content.

ration refers to the point in the data collection process when no new concept-relevant information is being elicited from individual interviews or focus groups, or no new information is elicited and deemed missing during cognitive interviewing. There is no fixed rule on either the sample size or the number of iterations needed to reach saturation. During the development process, evidence is based on empirical observation, where no new concepts or codes emerge after the *n*th interview or focus group [10,14]. Saturation can be evaluated and documented through a saturation table structured to show the elicitation of information by successive focus group or interview (individual or by set), organized by concept code. For practical purposes of budgeting projects, it is not uncommon to set a sample size of 20–30 interviews, even though saturation may occur earlier in the interview process. Saturation is then documented for where it occurs in the process, often during the interviewing process or sometimes at the end of all interviews. The risk of waiting to the end of the project before applying a coding frame and analyzing qualitative data is that saturation may not be reached and additional interviews will have to be conducted, or that saturation was evident at some point before the end of the designated stopping point, and the additional interviews were unnecessary.

It is not at all uncommon for participants in focus groups and interviews to stray from the intent of the interview, discussing issues or concepts important to them as individuals or as patients, but unrelated to the concept of interest (see Fig. 1, unrelated descriptions). These data are clearly not part of the core concept, are inconsistent with the intent of the instrument, and should be excluded from a description of the concept and saturation tables. This is a particularly important issue when analyzing qualitative data to evaluate the content validity of an existing instrument. For example, data from focus groups eliciting information on patient perception of anginal pain as part of an evaluation of an anginal pain scale may include comments from one or more patients describing their knee pain. These descriptions should be coded as part of the patient's comorbid experience, but excluded from the thematic summary and content item mapping.

Identifying and Resolving Threats to Content Validity

The fundamental question in the evaluation of a PRO instrument in the context of a labeling claim, whether an instrument is new, existing, adapted, or modified, is the adequacy of empirical evidence to support content validity for the desired claim. Four important threats to PRO content validity, in relative order of importance, are shown in Table 2. This table also provides an organizing heuristic for evaluating each threat; a discussion of this approach to resolving these threats follows:

Absent or Unclear Conceptual Match Evident between the PRO Instrument and the Intended Claim

As discussed earlier in the article, the conceptual match is the primary task in identifying the PRO instrument's conceptual framework, and specifying its linkage to the intended claim. If there is no clear match, then probably the most effective strategy is to identify another instrument to measure the claim with targeted concepts relevant for supporting the claim.

Lack of Direct Patient Input into PRO Item Content from the Target Population in which the Claim Is Desired

The patient population in which the PRO instrument was developed should be compared to the patient population targeted for enrollment in the clinical trials, to determine whether the instrument is appropriate for that population. This requires sponsors to carefully consider the inclusion and exclusion criteria of the clinical trial to identify important patient characteristics that may indicate a need for additional validation work.

Lack of Evidence Regarding Saturation—No Evidence that the Most Relevant and Important Item Content Is Contained in the Instrument

This threat addresses limitations in empirical data to confirm that the PRO item content associated with each concept captures the full range of important and relevant patient experiences across a representative sample of the targeted patient population. Evaluative interviews, focus groups, and cognitive debriefing interviews can be used to address this threat.

Modification of the Original PRO Instrument

Modifications to an instrument may include: changes in wording or content, changes in mode of administration, translation and cultural adaptation, and application to a different patient popu-

lation [10,15]. Two other ISPOR PRO task forces addressed best practices in relationship to changes in mode of administration [16], and translation and cultural adaptation [17]. Snyder et al. [15] note that modification of an existing instrument is an acceptable approach as long as the following conditions hold: the existing instrument has been adequately validated for a different application; items in the adapted instrument are relevant and appropriate for the target application; the instrument is able to be logistically utilized in the intended setting; and a new interpretation guideline is developed, as needed. The extent of documentation required to support existing instrument modifications will vary depending on the type of modification. If it has been modified in the way(s) described by Snyder et al. and is to be used in the regulatory approval process, Burke et al. [10] recommend that additional qualitative research be conducted.

Case Studies

To illustrate these considerations in evaluating PRO content validity, we present three case scenarios (Table 3). For each scenario, we identify the key question raised regarding content validity, identify the major threats, and offer sample approaches for remediation. These approaches are intended to be illustrative and in no way prescriptive.

In scenario A, a PRO measure was developed by clinicians to assess dyspnea in chronic obstructive pulmonary disease in clinical practice. The question here is whether the instrument could be used to evaluate treatment efficacy in clinical trials involving a new patient population, i.e., asthma. In this example, the first threat is that there may be aspects of the concept of "dyspnea" that are experienced uniquely by asthma patients and are not addressed in the instrument. A review of the literature and discussions with the developer may uncover qualitative study reports in which data from patients with asthma are presented as part of the instrument development process or by others interested in using the instrument in asthma. The review might also uncover independent qualitative studies examining the concept of dyspnea as experienced by patients with asthma with characteristics similar to those to be enrolled in the trials.

The second threat in this example is the potential that the instrument was developed without patient input. This is addressed by reviewing all of the documentation from the developer(s) for evidence of direct patient input. The material should provide a detailed description of all of the steps that were taken to identify instrument content, and to generate specific items based on patient data. Instruments developed before 2000 often lack this information entirely or in the detail needed to support labeling claims following the FDA Draft PRO Guidance [6]. Remediation for this problem is discussed below.

A third threat involves the change in intended use of the instrument, from clinical practice to clinical trial. From a content validity standpoint, the most important element of this threat is the method used to inform the overall design of the instrument, including recall period and item content. It is not uncommon for clinical instruments to be developed based solely on clinician expertise and experience, with content that addresses the specific information needs of the practice setting. For example, if the instrument was developed using qualitative research methods with direct input from patients with asthma similar to those to be enrolled in the clinical trial, and if results of this work are available, the magnitude of the threat declines. This is not the case in scenario A.

If documentation provided by the developer or available in the published literature is inadequate for informed decision-making, and/or insufficient for regulatory submission, additional work will be needed. In this scenario, one approach that could be used is to conduct focus groups and/or cognitive interviews with asthma patients who meet the inclusion/exclusion criteria for the product development trials, mapping the results to the instrument. This would enable the user to evaluate concept coverage and, if adequate, provide data for documenting the relationship between patient data and instrument content in the target population. If the content is found to be inadequate, this process would provide the sponsor with an opportunity to modify the instrument, with permission, and perhaps participation, of the instrument developer, with the potential for increasing the sensitivity of the instrument to detect treatment effects in clinical trials involving this new target population.

Scenario B illustrates a second example of a potential threat to content validity. In this example, an existing, multi-item, multiscale measure has been reduced to a shorter, more efficient assessment form using psychometric procedures, including cross-validation in the relevant patient samples, and has been shown to have nearly identical psychometric properties, including factor structure, to the original long form of the instrument. Here, the fact that there is evidence that the original factor structure is retained suggests that the original concept, or set of concepts, is still captured by the retained items in the short form. It is possible, however, that items representing concepts essential to the intended claim language were deleted in the process of creating the short form. For example, the physical function domain of a given instrument consists of multiple items that assess both upper and lower body function. During the item reduction process, all of the upper body functions are eliminated. The user must evaluate whether this change threatens the content validity of the remaining instrument in light of the items eliminated and the specific label claim language intended.

It is important to note that content validity involves the adequate sampling of content from the universe of all possible content to measure the concept of interest. To understand the extent to which content validity might be threatened by reducing certain items, cognitive debriefing interviews in the target population could be conducted to determine if item content considered essential to the concept had been eliminated. If this is the case, interviews will also determine the extent to which such item content was considered both relevant to the concept and important enough to the patients that the absence of the items may compromise measurement. Items that are redundant or closely related to other items provide no unique information and are unlikely to contribute meaningful information to the assessment of the concept. Eliminating redundant items is expected to leave the intended measurement concept intact. In this scenario, qualitative evaluation is recommended, even if results of quantitative assessment of the short form (reliability, validity, responsiveness) correlate with the long form.

In the last example, scenario C, the sponsor has selected an existing instrument that was developed to capture the claim-relevant concept(s) in the same target population as the intended study group, and appears to be appropriate for use in product efficacy trials. To be certain, the sponsor elects to verify content validity of the instrument in a small number of patients using focus groups. From these groups, the sponsor learns that the concept(s) is relevant, and the item content reflects the full range of relevant experience described by the patients. However, patients in the focus groups use different words or phrases than those used in the instrument. As discussed previously, items and response options comprising an instrument represent a consensus of the best wording for content validity. Cognitive debriefing interviews could be performed to verify that patients understand the instrument and interpret the items in a manner consistent

with the intent. If this is not the case, the original instrument may need modification.

Conclusions

Content validity refers to the extent to which an instrument contains the relevant and important aspects of the concept(s) it intends to measure. This article discussed the key issues involved in assessing and documenting the content validity of an existing instrument, including concept clarification, instrument identification, and initial review, as well as qualitative methods as they might be used to evaluate the suitability of one or more existing instruments. Case examples illustrate threats to content validity and various approaches for remediating these threats. Several tools were identified to aid in the evaluation of content validity, including end point models that describe the correspondence between concepts, measures, and labeling goals; the conceptual framework of the PRO instrument to evaluate and communicate the extent of the match between item content and targeted concepts; and qualitative research methods that form the empirical basis for evaluating and documenting content validity.

Source of financial support: The views expressed herein represent those of the authors and not those of Johnson & Johnson, the FDA, OLGA, United BioSource Corporation, University of Washington, or Pfizer.

References

1 American Educational Research Association APA, National Research Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: AERA, 1999.
2 Nunnally JC, Bernstein I. Psychometric Theory (2nd ed.). New York: McGraw-Hill Book Company, 1978.
3 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006;119:166.e7–16.
4 Frost MH, Reeve BB, Liepa AM, et al. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007;10(Suppl. 2):S94–105.
5 Haynes S, Richard DC, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol Assess 2005;7:238–47.
6 FDA. Food and Drug Administration draft guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register 2006.
7 FDA. Food and Drug Administration draft target product file—a strategic development process tool. Federal Register 2007.
8 Delasko J, Cocchetto DM, Burke LB. Target product profile: beginning drug development with the end in mind. UPDATE 2005;January/February.
9 Turner RR, Quittner AL, Parasuraman BM, et al. Patient-reported outcomes: instrument development and selection issues. Value Health 2007;10(Suppl. 2):S86–93.
10 Burke LB, Kennedy DL, Miskala PH, et al. The use of patient-reported outcome measures in the evaluation of medical products for regulatory approval. Clin Pharmacol Ther 2008;84:281–3.
11 Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 2007;10(Suppl. 2):S125–37.
12 Revicki DA, Gnanasakthy A, Weinfurt K. Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: the PRO Evidence Dossier. Qual Life Res 2007;16:717–23.
13 Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001;358:483–8.
14 Cohen D, Crabtree B. Qualitative Research Guidelines Project. Available from: http://www.qualres.org/index.html [Accessed December 23, 2008].
15 Snyder C, Watson ME., Jackson JD, et al. The Mayo/FDA patient-reported outcome instrument selection: designing a measurement strategy. Value Health 2007;10:S76–85.
16 Coons S, 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.
17 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 & linguistic validation good research practices task force report. Value Health 2009;12:430–40.