Establishing and Reporting Evidence of the Content Validity of Newly-Developed Patient-Reported Outcome (PRO) Instruments for Medical Product Evaluation: Good Research Practices

Donald L. Patrick PhD, MSPH, Laurie B. Burke RPh, MPH, Chad Gwaltney PhD, Nancy Kline Leidy PhD, Mona L. Martin RN, MPA, Lena Ring PhD

Part I
Developing Content for a New PRO Instrument

RUNNING TITLE: Developing Content for a New PRO Instrument
CORRESPONDING AUTHOR:
Donald L. Patrick PhD, MSPH, University of Washington, Box 359455 Seattle, Washington 98195-9455
donald@u.washington.edu

KEY WORDS: content validity; patient reported outcomes; FDA, EMA; quality of life
Authors are listed in alphabetical order by surname after the senior author. The views expressed herein represent those of the authors and not those of the University of Washington, Food and Drug Administration, PRO Consulting, United BioSource, Health Research Associates, AstraZeneca, or Uppsala University.
ABSTRACT

**Background:** A patient-reported outcome (PRO) instrument is a means to capture data for assessing treatment benefit or risk in medical product evaluation. Two articles in this issue present conclusions of an ISPOR taskforce convened to address good research practices for documenting content validity in newly developed PRO instruments. Content validity is the extent to which the content of a new PRO instrument adequately represents a given concept or set of concepts. We use the specific context of a PRO instrument newly developed to support PRO claims in medical product labeling. Paper I outlines steps for gathering and presenting qualitative evidence to support the inclusion of concepts in the new instrument. Paper II addresses how to gather evidence that persons in the target population understand the content of the new instrument. Both papers present suggestions for documenting the chosen qualitative theoretical approach, methods, results and conclusions. Adequate qualitative evidence is critical to ensure PRO instrument content validity. These papers do not address methods that mix qualitative and quantitative approaches to establishing content validity; however, the same qualitative research principles apply. Mixed qualitative and quantitative approaches to content validity testing will be addressed in future papers.

**Methods for Paper I:** Five good practices consistent with U.S. and European review processes are addressed in chronological order: 1.) plan context of measurement; 2.) develop protocol for qualitative concept elicitation; 3.) conduct concept elicitation interviews and/or focus groups; 4.) Analyze qualitative data for concept elicitation; and 5.) document concept development and elicitation. Illustrations are given of suggested ways to collect and present evidence.

**Results and Conclusions of Paper I:** Using qualitative evidence to support content validity requires a clear understanding of the actual intervention study design (e.g., the entry criteria for the clinical trial population) and the targeted context of measurement. Qualitative research applies to all PRO instruments used to support labeling claims and must be completed well before confirmatory (Phase III) trials are initiated to allow time for instrument finalization. The qualitative study protocols address a broad range of the target population demographics and characteristics and include a plan for data analyses. Conducting
Interviews or focus groups requires trained interviewers with appropriate quality controls. Qualitative analyses require trained coders, demonstration of saturation and clearly presented results supported by the transcripts of audio recordings. Detailed documentation of the entire concept elicitation process provides the body of evidence to support conclusions drawn by qualitative researchers that the instrument measures a certain concept. The evidence must support patients’ responses and outcomes data using the language of the instrument items correspond to the concept that is reflected by the instrument score(s) and that the concept is adequately covered by the instrument. The detailed documentation is also reviewed in a regulatory setting to determine whether medical product claims are truthful and not misleading when the instrument is used in an outcomes trial to measure treatment impact.
Background

The ISPOR Health Science Policy Council and the ISPOR Board of Directors recommended that an ISPOR Task Force be established on Good Practices in Establishing and Reporting Evidence of the Content Validity of Newly-Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation. The purpose of this task force was to extend the work of a previously published report on the use of existing or modified PRO instruments to support medical product labeling claims (1) by addressing methods for assuring and documenting the content validity of newly-developed PRO instruments.

The chair of this task force, (Donald L. Patrick, PhD) recruited members based on their experience as scientific leaders and practitioners 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: academia, government, research organization, and industry. In addition, forty-seven members of the ISPOR Patient Reported Outcomes Review Group provided written comments on the draft reports. In addition, oral feedback was provided at the PRO Forum held during the ISPOR 15th Annual International Meeting in Atlanta. The task force met regularly via conference calls and held one face-to-face meeting.

During content and outline development, the task force decided two papers would be needed: Part I covers the development of content for a new PRO instrument, i.e. concept identification to inform content and structure using qualitative focus group and interview methodology, while Part II covers item development and the assessment of patient understanding of the draft instrument using cognitive interviews and steps for instrument revision. The two parts are meant to be read together. Rather than prescriptive, they are intended to offer suggestions for good practices in planning, executing, and documenting the process of content validation of PRO instruments to be used in medical product evaluation.
PART I

Developing Content for a New PRO Instrument

Definition of Terms

The term “PRO” is often used interchangeably to refer to a concept, instrument, questionnaire, score, or claim. According to the FDA Guidance, a patient reported outcome (PRO) is any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else (2). In Europe, the European Medicines Agency Reflection Paper (3) released a reflections paper on the place of health-related quality of life (HRQL) in medical product development, specifying that this was one type of PRO. This task force report uses the term “PRO” to refer to the general concept or outcome of interest. “PRO” serves as the umbrella term covering all patient-reported outcomes, with HRQL one specific type (4). A PRO instrument or measure is a means to collect data. Questionnaires and diaries are examples of PRO instruments. The term instrument refers to item content (stem and response options), instructions, and recall period. PRO scores are numeric values or categorical assignment generated through the use of a PRO instrument and used to represent the PRO of interest.

In medical product development, PRO instruments may be used in clinical trials to capture and quantify treatment benefit or risk (5, 6), with the possibility that this information will be used to support a “claim” in medical product labeling. Within this context, it is useful to distinguish the PRO concept, claim, instrument, and score (5). For example, pain intensity is a PRO (the concept); decrease in pain intensity is a PRO claim; a 10-centimeter visual analog scale (VAS) assessing pain intensity, including the anchors, instructions, and recall period, is a PRO instrument; and the value a subject assigns to their pain intensity on the VAS is a PRO score.

Content validity is the extent to which the content of an instrument represents the most important aspects of a given concept (7), in this case, the extent to which it represents the PRO. In the FDA Guidance on PRO Measurement, content validity is defined by the empirical evidence showing that the
items and domains of an instrument are appropriate and comprehensive relative to its intended
measurement concept, population, and use (2).

Qualitative data are essential for establishing the content validity of a PRO instrument.
Quantitative data, including factor analysis and item response theory (IRT) analyses, may be supportive
but are insufficient on their own to document content validity for medical product development. This task
force report, Parts 1 and 2, summarize the elements of good research practices around the establishment
of content validity of a new instrument through qualitative research.

Good Practices in Eliciting Concepts for a New Patient-Reported Outcome Instrument

Table 1 lists five steps to elicit concepts for establishing and documenting content validity of a new
PRO instrument in a recommended chronological order, consistent with the Wheel and Spokes Diagram
contained in the Final FDA PRO Guidance (2). Within each step are recommended good research
practices.

Table 1 About Here

Good Practice 1: Plan the Context of Measurement

The development of an instrument, simple or complex, must start with a clear definition of the
concept to be measured in the proposed context of measurement. The purpose of Step 1 in Table 1 is to
ensure that the context is clearly defined and the approach for concept measurement is appropriate for the
intended context. In situations involving instrument development within a regulatory framework, context
considerations include the disease or condition of interest, target population, and treatment setting.
Consideration should also be given to the positioning of the measure in the hierarchy of clinical trial
endpoints. Clarification of the context of use and the role the measure will play in clinical trials informs
preliminary decisions on instrument scope of content, measurement structure, and mode of administration.
With this essential work established, qualitative research protocols can be developed to gather patient
input on the concept(s) of interest. Descriptions for each of the components of Step 1 follow below.
Disease models.

Development of a new PRO instrument for use in medical product evaluation often begins with a clear delineation of the concept of interest through the development of a disease model. Consideration is given to the pathophysiology and expression of the disease or condition, including its characteristic signs and symptoms in the target population. The relevant concepts measured by laboratory tests, performance assessments such as exercise stress or cognitive function tests, or standardized clinical observations are also identified. If symptoms are a defining characteristic, the appropriate symptom concepts cannot be determined based on a literature review and consultation with clinical experts alone. Qualitative research in the target patient population provides essential data on the patients’ perspectives of their symptoms. If the impact of health on related physiologic, psychologic, or sociologic concepts is of interest, those impact concepts may also be targeted for instrument development and require patient input.

Disease models help to clarify and focus the specific PRO concept of interest within the context of the entire disease process and the specific clinical trial population. Figure 1 illustrates a disease model for psoriasis with a proposed pathway linking risk factors, diagnosis, signs and symptoms, and impacts. The types of questions addressed by disease models include the following: Is the disease or condition characteristically symptomatic? Are these symptoms amenable to treatment? Are there functional effects of the condition, such as activity limitation due to symptoms that could be altered with treatment? What other outcomes might the treatment affect? What concepts should be the focus of efficacy evaluation? Because variability in measurement lowers the probability of detecting a meaningful treatment effect, the more specific the concept and the closer this concept is to the goals of the treatment, the greater the likelihood of success.

During the development of a disease model, consideration is given to the prevalence, severity, and characteristics of the condition, the treatment to be tested, the target population for treatment, and potential trial endpoints. Questions related to PRO candidates and trial design include many questions,
depending on the actual disease or condition and trial. Will patients enrolled in the trial be experiencing decrements in the symptoms, signs, or impacts that might be captured by a PRO instrument, so that the effect of treatment on this outcome can be appropriately tested? Will or can patients be screened for enrollment based on criteria specific to this outcome? In situations where the PRO is positioned as a secondary outcome and enrollment does not include criteria related to this secondary outcome, study results may be poor simply because a significant portion of study participants could not change with treatment.

Patterns of change over time in the PRO of interest are another consideration. Is the PRO relatively stable, with small changes over time? Is the condition acute with potentially large and/or rapid changes with treatment? Or is the condition chronic with an expectation for minimal or slow changes in the outcome of interest? For example, clinical trials for acute infectious conditions may be relatively short, while trials to demonstrate a survival advantage can involve relatively long observations. Trial design, including frequency of assessments, compliance, and missing data, is part of the context of use of a PRO instrument that will inform the content and structure, including the items, response options, and recall period.

**Endpoint models**

Endpoint models specify the primary and secondary endpoints to be tested in the target clinical trial(s). Example endpoint models from the FDA PRO Guidance were provided in Figures 1 and 2 the Guidance (2). Even when a medical product cannot be specified, e.g., in multi-sponsor instrument development consortia, the anticipated role of the instrument can be shown in one or more hypothetical or illustrative endpoint models to specify the context of use. In this case, the model(s) represent an educated prediction of the prioritization of study hypotheses in clinical trials in which the PRO instrument is to be used.

Of course, a new PRO instrument may serve as an exploratory endpoint in early trials, with the data used to test reliability, validity, and responsiveness. The endpoint models we are describing here pertain to future medical product development trials using current best clinical trial practices keeping in
mind that target patient populations may change to related severity or diagnostic groups. While it is important to be forward looking in developing a new instrument, this can result in a concept and instrument that is too generic, diluting measurement content, reducing reliability and sacrificing the near-term objectives.

*Literature review and experts.*

Disease and endpoint models both inform and are informed by existing knowledge or experience, published literature, and consultation with clinical content experts. Models focus the literature review and clarify the type of experts and the role they will play in the development process. Input from the literature and experts, in turn, are used to revise the disease and endpoint models as appropriate.

*Target population - cultural/language groups.*

As instrument development is planned, thought is given to the details of the target population, including the languages and cultures of patients likely to be enrolled in clinical trials. The extent to which the disease, standard of treatment, and measurement concept(s) are the same or differ across countries or cultures is considered. Literature and experts can help in this discussion. If the development program will be international and the concept is highly variable across countries, simultaneously developing an instrument internationally may strengthen and document cultural equivalence of the final instrument. If there is published or empirical evidence indicating concept stability across countries, it may be possible to develop the measure in one country with review by a PRO linguistic expert to facilitate ease of translation for future use.

*Preliminary decisions on the instrument content and structure*

As the context of use is identified and clarified, decisions are made concerning the optimal instrument structure and likely content. The following principles of good measurement are among those used during decision making: (1.) Consider both positive and negative content. For example, the effects of treatment may include positive effects on pain and negative effects on sleep. (2.) In general, respondents should not be asked to attribute the cause of their symptoms or experiences. It would be difficult, for example, for subjects to know whether their breathlessness was due to congestive heart failure
as opposed to other causes, such as aging, anxiety, infection, etc.  (3.) In general, respondents should not be asked to rate change over time, but rather should be asked to evaluate their current state with an appropriate recall period.  Change is then computed across evaluations.  (4.) Consider the method (self versus interviewer administered) or mode (paper-pen, electronic, voice response) of data collection early.  Switching methods or modes of administration between development and use may require an additional validation step to assure score equivalence.

Hypothesized Conceptual framework.

The considerations outlined above should lead to a list of the PROs of interest and the concepts and sub-concepts or domains comprising them.  The disease model shown in Figure 1, for example, shows two possible PROs of interest: psoriasis symptoms and Impacts.  Within each of these general PROs are concepts and sub-concepts suggestive of instrument content, e.g., pain, itching, burning etc.  This information informs the development of the qualitative elicitation protocol and the interview or focus group discussion guide.  As outlined below, the guide includes reference to what the interviewers might expect to hear and areas requiring greater clarity, with the understanding that new information may be uncovered, contributing to the conceptual focus and accuracy of the instrument.

An example of a conceptual framework for a PRO evaluating the concept of pain is shown in Figure 2.  Note that the category of pain quality is divided into deep pain and surface pain.  These two concepts are further divided into aspects of pain quality.  This conceptual framework will help with the coding dictionary developed later in the process.

Good Practice 2:  Develop the research protocol for qualitative concept elicitation

The study protocol and interview guide provide documentation of the pre-specified plan for identifying the sample, conducting interviews or focus groups, and analyzing data that will inform the content and structure of the new instrument.  Contents of the study protocol include:  Study sample, data collection method, setting, materials and methods, and analyses.
Study Sample

Demographic and clinical characteristics of the sample should match the target population, i.e., the intended clinical trial sample. For example, if clinical trials will include patients who have either psoriatic arthritis or plaque psoriasis, both types of patients are included in the qualitative study sample in order to allow for the full range of comments and expressions to arise. Clinical sites and methods for participant recruitment should be selected with this goal in mind. When evaluating clinical sites and/or locations for possible participation, considerations should be given to geographic, educational, ethnic and racial diversity, and the availability of clinical information needed to characterize and evaluate sample characteristics in the final report.

Estimating the sample size for a qualitative study can be challenging. In quantitative research protocols, sample size is estimated using analytical techniques requiring projections of magnitude likely to be observed in the study (e.g., means, differences, variances, proportions, confidence intervals) together with the desired power and a significance criterion. In qualitative research, sample size estimation is based on projections of the data needed to reach “saturation”. Discussed further in Section 4, saturation is “the point at which no new concepts [relevant to the concept of interest] are forthcoming from the population being interviewed” (8) When the concept of interest is clearly defined and relatively narrow in scope and the target population is largely homogenous, relatively few participants (e.g., 15 to 20) may be required to achieve saturation. In contrast, situations involving a very broad, poorly defined, or multidimensional concept or heterogeneous target populations will involve larger sample sizes (e.g., 40 or more). As noted in the FDA Guidance (2), “the number of patients is not as critical as interview quality and patient diversity included in the sample in relation to intended clinical trial population characteristics.”

Data collection method

Individual interviews and focus groups are the data collection methods used in qualitative research involving concept elicitation for instrument development purposes (9). A summary of the advantages and disadvantages these methods are shown in Table 2. Focus groups are economical and can stimulate discussion of topics and comparison of experiences across participants that cannot be captured in
individual interviews (9-12). Unfortunately, there are also risks associated with focus groups, particularly when run by inexperienced or untrained leaders. One example is a highly vocal, assertive participant who dominates or leads the discussion, minimizing participation of other group members and resulting in content, tone or perspectives that do not necessarily represent those of individuals or the group as a whole. Interviews are ideal for concepts that are sensitive or target populations/people unlikely to volunteer or share information in a group setting (13). There are also disadvantages to individual interviews. For example, by design, interviews must be conducted sequentially or by multiple interviewers, both of which are more expensive and time consuming (14). Interviews are usually the best methodology for concepts that are sensitive or target populations unlikely to volunteer or share information in a group setting.

Table 2 About Here

Setting

Focus groups and interviews may be conducted in in-patient settings, out-patient clinics, or dedicated research facilities. Interviews may also be conducted at participant homes or, in some cases over the telephone (e.g., rare, episodic, or contagious conditions). The appropriate setting depends on the target population, including illness severity or contagiousness, physical mobility, psychological state, or other factors that would affect a person's ability to travel or participate. The setting should optimize the extent to which the sample is consistent with the target population by making participation accessible.

Materials and Procedures

The interview or focus group guide includes the questions that should be addressed and how the interviews or focus group should unfold for optimal clarity and data quality. It is not a script to read verbatim, but is a manual that provides the interviewer with an organized summary of the topics to be discussed, specific questions for each topic, and sample probes that can be used to further explore areas when needed. Exploratory questions may be included to uncover features of the condition or its treatment that may not be well understood through previous research and clinical experience.

The specific content of questions comprising the interview guide is dictated by the context of
measurement, including the disease and endpoint models and draft conceptual framework. For example, if pain is hypothesized as an important symptom in the disease model, the interview/focus group guide includes questions to understand the patient’s experience of pain, which may include frequency, severity, duration, and or impact. The reference timeframe, that is the timeframe the participants are asked to consider as they respond to the questions, will also depend on the PRO and measurement context. For example, when developing a measure for chronic heart failure patients, participants may be asked to recall and describe a recent acute episode or hospitalization or their experiences during the current day or week. In general, it is desirable for the reference timeframe to be as close as possible to the interview or focus group, in order to diminish recall errors and bias. One method known as the day-reconstruction approach (15) can be used to focus a participant on a specific day as they describe symptoms, impacts or other experiences relevant to the target concept.

Unless carefully worded and conducted, interview questions and procedures can introduce bias into the data. For example, certain closed-ended or highly specific questions can be leading, such as “you experienced pain in your knee today, right?” or “how depressed were you during this event?” Questions should be open-ended whenever possible and worded to encourage spontaneous information from the participant without pointing them toward a specific response. With this in mind, open-ended questions that are too broad can be confusing to participants. “What was yesterday like for you?” or “Tell me about your condition”, for example, lacks the specificity required for participants to address the concept of interest and can lead to irrelevant data. Open-ended question should include parameters consistent with the concept of interest. If the concept of interest is knee pain, the interviewer could ask: How did your knee feel yesterday? With probes to better understand the nature and characteristics of the experience offered by the participant. This approach provides data on the words and phrases participants use to describe their condition that will inform instrument content.

Interview questions can also address multiple dimensions of a concept. For example, it may be useful to understand the severity, duration, and frequency of a particular symptom. Pain that can be severe, but doesn’t last long or occur very often may be a very different experience from pain that is
moderately severe, but occurs frequently and lasts for a long period of time. Understanding these dimensions of an experience can be useful for developing a new instrument based on a complete picture of a participant’s experience. The following list of questions show how more specific symptom-related information might be obtained once the symptom has been elicited by more open ended questioning: How often do you have (symptom X)? How severe is the (symptom X)? How long does it usually last? Does anything make (symptom X) better or worse? Please tell me more about that. Do you have any other sensations or symptoms when you feel (symptom x)? Questions to elicit information about symptom impact might include: How do your symptoms affect or influence your everyday life? Probes might include: How does symptom X affect your daily activities? Does it affect your relationship with others? Tell me more about (the difficulty you have performing activity X).

Once a draft interview guide has been created, it is reviewed by other qualitative researchers for possible difficulties in flow, redundancy, poorly formulated questions, and the appropriate use of terminology and probes. The draft guide should be pretested with study naïve individuals or colleagues or, ideally, pilot tested in the target population to identify areas that do not flow easily or may confuse respondents before primary data collection begins.

Analyses

As with a clinical trial, the interview protocol should also include a plan for analyzing, summarizing, and interpreting the interview data. Unlike quantitative analyses, there are no inferential statistical tests involved. Rather, this portion of the protocol describes the methods that will be used to identify, code, and summarizing themes, procedures for quality control, and methods for determining and documenting saturation. Qualitative analyses are discussed further in section 4.

Good Practice 3: Conduct the concept elicitation interviews and focus groups

The research protocol must be reviewed and approved by an appropriate institutional review board prior to the initiation of subject recruitment and data collection. Sites are provided a copy of the study
protocol and trained on inclusion/exclusion criteria, sample monitoring, recruitment processes, and informed consent procedures.

Interviewers and focus group facilitators should be experienced in qualitative research methods and trained on the background and objectives of the protocol. Mock interviews or focus groups may be used to help the interviewers/facilitators develop a complete understanding of the questions and process and assure a smooth, clear data collection process. Sustained interaction with interviewers is important to establishing and maintains quality of data collection.

Core competencies in concept elicitation interviewing are shown in Table 3. The concept elicitation process is intentionally broad in order to explore and define information from the perspective of the patient. A well constructed interview guide defines the broad territory of discussion, leaving no need for the interviewer to censure or discount participant responses. Although discipline is needed to keep the participant or focus group “on task”, interviewers should avoid being overzealous in assuming irrelevance, favoring an open dialogue among participants to encourage participation. Interviewers should be aware that their body language and actions, such as nodding in agreement, frowning, or sighing, can communicate approval or disapproval of the participant’s contribution, altering the content or emphasis of subsequent information. Interviewers should remain neutral, while conveying genuine interest to encourage open and honest communication. Hallmarks of interviewer skill rest on the ability to get the participant to talk about the areas and topics of interest in a natural conversational engagement, where they feel they are being heard and respected.

Table 3 About Here

Concept elicitation interviews and focus groups are recorded (either audio or video) to fully capture the context and content and produce transcripts that form the data for analysis. Audio recordings are generally preferred because they are easier to perform and transcribe, facilitate participant anonymity, and are generally more comfortable for participants, particularly when sensitive topics are being discussed. Regardless of recording method, participants need to be assured of the confidentiality and limited usage of the recorded materials from their interviews. In addition to being essential for data analyses, recordings
can be monitored for quality assurance by a senior interviewer who provides feedback to the interviewer to
maintain or improve the quality of data collection throughout the duration of the study by improving
question clarity, altering probes, and/or pursuing specific aspects in greater detail.

Recording frees the interviewer or moderator from note taking in order to engage fully with the
participant(s). For focus groups, an assistant moderator is often useful to observe the group and take
notes to facilitate data interpretation. These notes include a seating chart with participant initials and key
points associated these initials. This also helps in checking the transcriptions of focus group recordings.

Transcriptions of the audio/video recordings need to be verbatim and reviewed, quality checked,
and cleaned by the facilitators/interviewers and associates. Cleaning includes: (a) removal of any
personal identifiers; (b) correction of any medical terms that the transcribers did not recognize or
misspelled; and (c) removal of any clearly extraneous narrative (for example, the participant answers their
cell phone or the nurse walks in with a message). Dialogue that is related but not central to the purpose of
the interview can be retained in the transcript and separated during the coding process to document the
irrelevance of the information for data analyses. Transcript quality is assessed through the direct
comparison of voice and transcript files, generally performed randomly. Once transcripts have been
quality checked and cleaned, qualitative analysis begins.

**Good Practice 4: Analyze the Data**

Analyze according to the theoretical approach. There are multiple theoretical approaches and
methodologies that can be applied to qualitative research procedures and data analyses, including
phenomenology, grounded theory, content analysis and thematic analysis(16-19). In qualitative research
to inform instrument development, data collection and analyses are interrelated and concurrent rather than
linear: “Analysis is the interplay between researcher and data. It is both science and art” (20). All of these
approaches are idiographic (focus on the individual) in contrast to a quantitative nomothetic paradigm
(focus on the general) founded in positivism (21). Across all qualitative methods, the purpose is to
understand participant perspectives and experiences using “decontextualisation” i.e., assigning of codes and “recontextualisation” (i.e., reducing data around central themes).

Phenomenology as an overarching theoretical framework and grounded theory as a specific methodology have been proposed as most appropriate for the development of a new PRO instrument(18). An adaptation of grounded theory has also been proposed (14) which allows for the use of prior knowledge in the analysis of data. This added deductive element to an otherwise inductive approach is consistent with the need to draw from existing information, pulled together as part of context-of-measurement development (Step 1), to identify themes and concepts in the data and interpret the results in light of the ultimate goal, to develop a new PRO instrument for a specific use. This approach also permits moving back and forth between a hypothetico-deductive and inductive approach where the developer’s understanding can change based on new information and / or observations, resulting in an iterative process of instrument development.

It is important to clearly describe how the data were analyzed, i.e., what was done and why. Existing guidelines for performing qualitative research can aid in structuring the description, evaluating the process used and, and determining how best to present and discuss results (22-24),(25, 26).

Coding qualitative data for instrument development

The primary goal of transcript coding is to organize and catalog a participant’s descriptions of their experiences within the context of measurement. The coding process for different qualitative approaches share methodologies for decontextualisation and recontextualisation, even when the coding focus differs. For example, based on a phenomenological approach, one can identify descriptions of the phenomenon that are universal (phenomenology); based on grounded theory, one can use open coding (examining, comparing, conceptualizing, and categorizing data), axial coding (reassembling data into groupings based on relationships and patterns within and among the categories identified in the data); and selective coding (identifying and describing the central phenomenon, or “core category.
The “coding framework” is an initial structure or organization of codes for grouping clusters of information that form a coherent theoretical unit. This framework is based on the disease model and draft conceptual framework developed at the onset of the work. A preliminary coding framework is developed and revised during data analyses based on information and insight gained during data review, including the development of new codes to represent clusters of new information. Data coding is an iterative process, and should include opportunities to be re-examined and re-analyzed until no new codes are identified and all relevant concepts have been assigned one or more codes.

Figure 3 shows the various inputs into the development of a coding framework (structure to hold codes) to the completed “coding dictionary” (document inclusive of all codes assigned with definitions as appropriate for standardization, clarity, and communication). A coding framework provides patient-based insight into the relevance of concepts included in the disease model and conceptual framework. A coding dictionary is used to assure consistency in coding across data analysts or coders and to document and communicate the meaning of the codes to external reviewers.

Presentation of the coded qualitative data is intended to identify both the predominance of participants expressing the concepts and to provide a description of the language that the participants use to talk about those concepts. Depending on the qualitative approach, the presentation of codes and themes might differ. A thematic ‘map’, i.e., an overall conceptualization of the data patterns and the relationships between them will be produced when using a thematic analysis(19).

Computer-assisted qualitative data analysis software programs, such as Atlas.ti (27) can be used to organize the data and coding scheme for easier retrieval and analyses. These programs do not assign codes to the data; skilled decision making is still needed to allocate participant expression of concepts to the appropriate code.

Assessing Saturation.

Best practice is to code and assess saturation at multiple points during the data collection process, Data should be transcribed and coded on a rolling basis with regular intervals of assessment to evaluate
the consistency of the code assignment process, adequacy of the coding framework and to monitor the appearance and organization of newly appearing concept codes. Careful monitoring during the coding process and a phased approach to assessing saturation provides the researcher with insight into the data as the study progresses and an opportunity to return to the field for comprehensiveness or clarity.

To assess saturation of concept, transcripts and coding can be evaluated after a set of 5 to 8 interview or focus group transcripts become available. A saturation table is used to track either the new appearance of concepts, or noting all occurrences of the concept across the transcript groups. Data are examined for either the continued identification of new concepts (newly appearing codes) or codes requiring further examination to confirm relevance or the attainment of saturation.

Codes are identified in each next set of transcripts and compared with the codes that appeared in the previous groups. In the best case scenario, saturation is documented by showing no new concepts arising in the last several interviews or final focus group. In reality, it is not uncommon for a new concept to arise late in data collection process. Scientific judgment, including knowledge of the field and consultation with experts, is used to determine if this new concept is an outlier, i.e., reflecting a relevant but unusual case, and further judgment is required to determine if additional data collection is required or warranted to re-assess saturation following this late revelation.

Multiple coders

Best practice in analyses of qualitative data from elicitation interviews involves two or more coders. Each coder is carefully trained around the purpose of the study, target concept, nature of data itself, the coding framework, and the coding dictionary. Each coder completes 1-2 transcripts and meets to compare codes assigned, identify areas of consistency and inconsistency, reconcile the codes on these transcripts, and revise the coding framework and dictionary for clarity and to enhance consistency in subsequent transcript coding. This process is repeated regularly throughout the coding process. An agreement is defined as a set of words or phrases identified as reflecting the same code and/or sub-code. Given the nature of qualitative data, flexibility is permitted around the words that constitute the word set or phrases. For example, two coders assigning the codes “pain” and “pain with kneeling” to the transcript
text “You know I am always in pain when I kneel” would be considered in agreement, even though one is more specific than the other.

Assuring coding precision can take several forms. One approach is to have a “super coder” review all data to assure consistency across coders. A second approach is to draw a random selection of transcripts that are dually coded and assess inter rater agreement. Through discussion of coding and reconciliation when disagreement between coders is uncovered, greater than 90% agreement can be reached. These methods are similar to those in interviewer coded audio recordings using psychiatric ratings scales where inter-rater agreement is critical and inter-rater agreement is assessed until it reaches 90% or higher (28). Regardless of the approach used, the coding method and procedures for quality assurance should be carefully documented.

Multi-Vectored Analysis of Qualitative Data

Analyzing qualitative data is a multi-vectored assessment where different vectors of information are gained throughout the qualitative interview process. These often include: pre-specified concepts (symptoms, signs, limitations, worries, impacts, etc.); concepts participants report spontaneously versus those they recognize when probed; predominant language participants use to express various concepts; variability in experience around concepts; the most meaningful way to address concepts (attributes of frequency, severity, duration); and the degree of difficulty, bother and/or impact.

The selection of any one format, focus or analytic approach is dependent on the purpose of the study. For example, an exploratory analysis aiming to elicit concepts for theory development might focus more on presenting information vectors like “relevant concepts” and “patient language”. In contrast, analyses for instrument development requires a focus on information vectors to successfully craft items, response options, instructions and recall, such as the “attributes” and “variability” associated with the target concept. Each information vector can have one or more uses and can be presented for assessment in a number of formats (i.e., by content, by predominance, by actual scores, or by proportion), depending on the type of inference to be drawn and the framework and analytical method chosen. This multi-vectored approach is illustrated in Figure 4.
The language in participant quotes provides a rich picture of the participants’ experiences with the target concept. In qualitative research for instrument development, the goal is to understand, organize, and communicate the meaning of the data and translate that meaning into a quantitative measure. The analysis of qualitative data is not quantitative; there is no effect size, significance level, or other quantitative metric. The goal of qualitative analyses to understand and communicate the meaning embedded in a dataset comprised of words and phrases. This is done by analyzing, organizing and summarizing the data in a manner that shows the relationship between the concepts, the words and phrases, and the final PRO instrument. Because each vector of information involving patient input contributes a unique aspect of understanding and communication, the use of multiple vectors of information provides an instrument developer with greater confidence that the concept is understood and the instrument adequately expresses this understanding.

Good Practice 5: Document Concept Development and Elicitation

The FDA PRO Guidance lists the information to be provided by Sponsors in PRO dossiers in an Appendix (2)pp 35-39). The FDA Guidance proposes an order and taxonomy based on the wheel and spokes diagram that provides a logical flow for organizing the report to support the PRO being submitted for review in relation to the claims desired and the development process ((2), page 7). For both the FDA and EMA reviews, documentation begins with the PRO instrument to be reviewed, followed by a description of the steps used to identify concepts and create the instrument.

Concept elicitation methods are part of the evidence supporting content validity as recommended in the first two spokes of the FDA diagram. Essential documentation of content validity includes both concept elicitation discussed in this paper and cognitive interviewing discussed in the next (Part II). This qualitative evidence may be accompanied by supplementary quantitative evidence that confirms or revises the proposed conceptual framework. Essentially the early content validity documentation provides evidence that the proposed instrument captures the most important concepts as viewed by the target
population, and that the concepts are complete and relevant to persons in the target population. This evidence is specific to the planned clinical trial population and indication, i.e., the context of measurement. Consistent with the FDA Guidance, documentation of the concept elicitation phase of instrument development include the following elements:

- Target claims and description of the target population (i.e., from Target Product Profile)
- The preliminary and final disease model
- The underlying endpoint model
- Preliminary and revised conceptual framework for the PRO instrument based on qualitative studies conducted prior to testing of measurement properties
- Literature review and documentation of expert input
- Qualitative study methods and results, including protocols, interview guides, and results
- Evidence of saturation.
- Origin and derivation of concepts captured in the PRO instrument
- Summary of qualitative data supporting the concepts, items, response options, and recall period

Organizing the document in a manner consistent with recommendations contained in the FDA PRO Guidance makes it easier for reviewers to determine if the essential elements of qualitative development of a new PRO instrument are included in a submitted dossier. Further recommendations on documentation of item wording, cognitive interviewing and the final item tracking matrix prior to quantitative evaluation are contained in the following manuscript.

Conclusion

This paper outlines the steps needed to derive a new PRO instrument for use in medical product development trials evaluating the benefits and risks of treatment. The paper covered the steps of concept elicitation, from determining, defining and documenting the context of measurement to the analyses of qualitative data from interviews and focus groups and the documentation of methods and results of this work. Examples have been provided to clarify specific steps and inform the development of
documentation needed to support the content validity of the new measure. Paper II of this 2-part task force report covers the creation of the new PRO instrument, evaluating its clarity and content validity through cognitive interviewing, and documenting this work for medical product evaluation.
REFERENCES


Table 1  Five Steps to Elicit Concepts for New Patient-Reported Outcome Instruments and Document Content Validity Consistent with Good Research Practices*

1. **Determine the context of measurement**
   - Develop a hypothesized disease model based on literature, experts, and patients
   - Name and define the concept within the context of a clinical trial end-point model
   - Select and define the target population
   - Conduct a literature review, prepare list of candidate items from disease model and existing instruments addressing the same concept, and consult content experts
   - Select the target cultural/language groups
   - Make preliminary decisions on instrument content and structure
   - Develop an hypothesized conceptual framework for the instrument

2. **Develop the research protocol for qualitative concept elicitation**
   - Define the target sample characteristics
   - Select the data collection method - focus groups, individual interviews, both
   - Determine the setting and location for data collection
   - Develop the interview guide – Draft, pilot, revise
   - Determine quality control procedures for data collection and monitoring
   - Develop a preliminary qualitative analysis plan

3. **Conduct the concept elicitation interviews and focus groups**
   - Obtain IRB approval
   - Recruit and train sites
   - Recruit participants; monitor sample characteristics to assure representation
   - Select and train interviewers
   - Conduct interviews – implement quality control measures
   - Record or videotape interviews
   - Transcribe and clean transcripts

4. **Analyze qualitative data**
   - Analyze qualitative data according to theoretical approach used
   - Establish preliminary coding framework; update as data are coded
   - Establish coding procedures and train coders
   - Organize data using a qualitative research software program Assess saturation
   - Interpret results
5. Document concept development and elicitation methodology and results

- Provide context for use
- Specify and define the concept
- Denote the target claims and population
- Provide a disease model and an endpoint model

- Provide supporting documentation for concept
- Show the original and revised conceptual framework
- Summarize the literature review
- Document input from content experts
- Present the methods and results of qualitative research
- Provide clear evidence of saturation

*Steps to develop an instrument, evaluate the new measure through cognitive interviewing, and document that aspect of content validity are addressed in Part 2 of the Task Force report.
## Table 2  Focus Groups & Interviews: Advantages and Disadvantages

<table>
<thead>
<tr>
<th></th>
<th><strong>Focus Groups</strong></th>
<th><strong>Interviews</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages</strong></td>
<td>▪ Rich source of data</td>
<td>▪ Get more in-depth and detailed information about an individual's experience</td>
</tr>
<tr>
<td></td>
<td>▪ Allows individuals to use ideas of others as cues to express their own views</td>
<td>▪ Can be useful for sensitive topics</td>
</tr>
<tr>
<td></td>
<td>▪ Participants can compare their experiences with others</td>
<td>▪ Data can be easier to analyze</td>
</tr>
<tr>
<td></td>
<td>▪ Able to reach many participants at once</td>
<td>▪ Scheduling can be easier</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>▪ Data can be tough to analyze because talking can be in reaction to the comments</td>
<td>▪ It may take longer to collect the data</td>
</tr>
<tr>
<td></td>
<td>of other group members</td>
<td>▪ Limited to one participant's view at a time; no peer comparison</td>
</tr>
<tr>
<td></td>
<td>▪ Moderators need to be highly trained and able to lead the group</td>
<td>▪ Interviewers need to be trained with excellent one-on-one communication skills</td>
</tr>
<tr>
<td></td>
<td>▪ One strong group member can sway tone of entire group</td>
<td>▪ May be more costly (e.g. travel, room rental, transcription fees, etc...)</td>
</tr>
</tbody>
</table>
### Table 3  Form for Evaluating Core Competencies in concept elicitation interviewing

<table>
<thead>
<tr>
<th>FOCUS OF EVALUATION</th>
<th>Criteria met?</th>
<th>Issues Found In Evaluation</th>
<th>REMEDIES NEEDED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PREPARATION to start interview w/subject</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepared for Interview, Familiar with interview content, mechanics of worksheets, tallies, record keeping</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demonstrates understanding of interview content, comprehends patient response in context of goals</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PROTOCOLS followed as identified in interview guide</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identified primary purpose of Interview/Focus Group to subject(s)</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adhered to interview guide</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covered all probe content</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allowed time for participant to spontaneously respond to probes before offering examples</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thoroughly explored responses to probes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asked for additional comments at completion of interview</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>COMPETENCIES Demonstrated in Conduct of Interview</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsive to lack of understanding of subject to question/topic – able to reframe question for participant understanding</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allowed subject time to respond without interrupting/rushing subject</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offered a minimum of 3 examples where needed</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control of interview (keeping subject on topic, familiar with interview logistics)</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stayed neutral – avoids confirming</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOCUS OF EVALUATION</td>
<td>Criteria met?</td>
<td>Issues Found In Evaluation</td>
<td>REMEDIES NEEDED</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>subjects responses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actively promoted in depth responses from subject</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avoiding leading Questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using participant language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognizing the symptom has already been explored – spontaneously by participant</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GENERAL COMMENTS**

Overall feeling from the interview:

- **Things done well:**
- **Things that need improvement:**

Printed with permission of M. L. Martin, Health Research Associates
Figure 1. Disease Model for Psoriasis to Aid Development of Concepts to Measure in Context of Disease, Population, and Treatment
Figure 2: Example conceptual framework for a PRO evaluating the concept of pain quality
Figure 3: From coding framework to coding dictionary

CODING FRAMEWORK
- Literature
- Clinical Experts
- Interview Guide
- Hypothesized TPP
- PRO Instrument Review
- Hypothesized Conceptual Framework

CODING DICTIONARY
- Contains all codes assigned, grouped by concept, and
- Patient Interview results
- All Code Assignments from transcripts

Printed with permission of M. L. Martin, Health Research Associates
Figure 4: Multi-Vectored Approach to Understanding Qualitative Data

- **Patient Language**
- **Relevance of Concept**
- **Attribute to Measure**
- **Necessary Coverage (degree of importance to patient)**
- **Variability**
- **Degree of Difficulty**
- **Degree of Bother**

**INTERPRETATION & MEANING**
- Presenting the relation of qualitative results to concepts
- wording, with most meaningful and appropriate measurement design, response options and recall period

**Essential Coverage (degree of importance clinically and to measurement strategy)**