Qual Life Res (2010) 19:1087–1096
DOI 10.1007/s11136-010-9677-6

PRO development: rigorous qualitative research as the crucial foundation

Kathryn Eilene Lasch · Patrick Marquis · Marc Vigneux · Linda Abetz · Benoit Arnould · Martha Bayliss · Bruce Crawford · Kathleen Rosa

Accepted: 6 May 2010 / Published online: 30 May 2010

Abstract Recently published articles have described criteria to assess qualitative research in the health field in general, but very few articles have delineated qualitative methods to be used in the development of Patient-Reported Outcomes (PROs). In fact, how PROs are developed with subject input through focus groups and interviews has been given relatively short shrift in the PRO literature when compared to the plethora of quantitative articles on the psychometric properties of PROs. If documented at all, most PRO validation articles give little for the reader to evaluate the content validity of the measures and the credibility and trustworthiness of the methods used to develop them. Increasingly, however, scientists and authorities want to be assured that PRO items and scales have meaning and relevance to subjects. This article was developed by an international, interdisciplinary group of psychologists, psychometricians, regulatory experts, a physician, and a sociologist. It presents rigorous and appropriate qualitative research methods for developing PROs with content validity. The approach described combines an overarching phenomenological theoretical framework with grounded theory data collection and analysis methods to yield PRO items and scales that have content validity.

K. E. Lasch (✉) · P. Marquis · M. Bayliss · K. Rosa
Mapi Values, Boston, MA, USA
e-mail: kathy.lasch@mapivalues.com

M. Vigneux · B. Arnould
Mapi Values, Lyon, France

L. Abetz
Mapi Values, Bollington, UK

B. Crawford
Mapi Values, Tokyo, Japan

Keywords PRO development · Qualitative research · Grounded theory methods

Qualitative research in PRO development

Patient-reported outcomes (PROs) in clinical trials, effectiveness studies, and public health research have been defined as "any report coming directly from subjects without interpretation of the physician or others about how they function overall or feel in relation to a condition and its therapy" [1, p. 125]. The value of qualitative research in the development of PRO measures has been recognized for many years. Witness the growing acceptance of such research by a new edition of a book that devoted a brief chapter to qualitative research in an otherwise comprehensive volume on quantitative methods that are used to measure quality of life in clinical trials [2]. A more recent focus has been placed on the concepts being measured and their meaning—not in terms of correlation coefficients or factorial structure, but their authenticity for subjects, i.e., their content validity. The emergence of content validity as a construct was to guard against strictly numerical evaluation of tests and other measures that overlooked serious threats to the validity of inferences derived from their scores [3]. This article presents an approach incorporating an over-arching phenomenological approach into grounded theory data collection and analysis methods to most accurately include the subject's voice in PRO development.

The quest for authenticity in instrument development evolved from a pragmatic approach ranging from literature review, clinician expertise, and the psychometric performance of items from large samples and batteries (e.g., Medical Outcomes Study Short Form [SF-36]) to direct involvement by subjects in item generation [4]. When

subjects have been included to date, however, the systematic analysis of their words and the link from their words to concepts underlying items is usually neither documented nor transparent. Transparency and systematization, however, are considered hallmarks of good qualitative research [5]. Their absence in qualitative research in the PRO field makes it difficult to communicate and compare results. Other essential issues in the conduct of rigorous qualitative research for PRO development include: who does one interview, how does one analyze the data systematically and transparently, how does one develop a conceptual framework to undergird a questionnaire from participants' responses, and, above all, what overarching theoretical framework (a guide as to which concepts and which relationships between those concepts should be the focus of a research study), if any, would best serve PRO development? A conceptual framework as defined by the Food and Drug Administration (FDA), one of the major constituents for PROs, represents the demonstrated relationships between and among items on a questionnaire and domains (multidimensional concept in which items are grouped together)[1, 6].

The FDA issued a PRO draft guidance document in 2006 and a final Guidance to Industry in 2009 that, when followed, makes it critical for instrument developers or reviewers to use and understand state-of-the-art methods in qualitative research [6–8]. Adherence to this guideline is required if the questionnaire is intended as an endpoint to evaluate treatment benefit assessing clear concepts that might support a labeling and/or advertising claim. Recently, members of the Study Endpoints & Label Development (SEALD) division in the FDA gave presentations at the 45th Annual Meeting of the Drug Information Association in which they emphasized the importance of content validity as an important and qualitative measure. Content validity, in general, means that a measure captures what it intended to measure. In these presentations, the FDA more specifically defined content validity of a PRO as (1) evidence that the items and domains measure the intended concepts, as depicted in the conceptual framework and desired claim; (2) evidence that the items, domains, and concepts were developed with subject input and are appropriate, comprehensive, and interpretable by subjects; and (3) that the study sample is representative of the target population.

Both the collection of qualitative data and its analysis have become more systematized and rigorous in the past 30 years as health researchers have increasingly incorporated them into their work. The most informative ways to interview participants have been refined. Even when provided with discussion guides and training to conduct focus groups or in-depth interviews, however, interviewers untrained in qualitative research methods use these guides as though they were conducting a structured interview. They often ask questions that put words in the subjects' mouths and do not dig deeper than what is directly asked (*or* rarely go beyond the scope of questioning). Probes that ask a study participant to describe more fully the meaning of a concept that is spontaneously offered are rarely used. Using guides as rigid scripts limits the collection of data that is ideal for capturing subjects' meaning of the experience of a condition and its treatment. In addition, the PRO field generally has not taken full advantage of the decades of knowledge in the field of survey research psychology to construct items and responses that most clearly depict the experience of a symptom or an impact of a treatment or a condition [9–11].

Researchers have published or presented criteria on how to evaluate qualitative research in health literature in general and in the development of PROs in particular [12–16]. However, very little information is available in the PRO field on how to collect and analyze qualitative data compared to the plethora of literature on psychometric methods to support the validity of PROs. Only one article to our knowledge, published in 2008, specifically discusses qualitative research methods to assure clarity and content validity in PROs [17].

We present an approach to develop a PRO instrument with content validity. This approach was developed by an international, interdisciplinary team of psychologists, psychometricians, regulatory experts, a physician, and a sociologist with over 25 years of experience conducting qualitative research. We describe how qualitative research and the psychology of survey response may best be applied to capture both the meaning of medical conditions to subjects and treatment impact.

Brief background: psychology of survey research and qualitative health research

Similar to its quantitative equivalent, qualitative research is an umbrella term for various theoretical models and data collection methods [18, 19]. Anthropologists, sociologists, nursing researchers, and, recently, psychologists have applied various methods and theories to the health arena [7, 20–23]. There is also extensive literature on the psychology of survey research that addresses how respondents answer items on a questionnaire [9–11, 24–28]. The most commonly used cognitive model is the question/answer model proposed by Tourrangeau in 1984 [29]. This model identifies the cognitive stages in answering a survey question, including comprehension, retrieval, judgment, response selection, and response reporting [25]. This literature takes into account the interactive aspects of the interview context and the cognitive processes that are involved in answering items. Its focus has been on the improvement of

questionnaire design rather than the blank slate involvement of subjects to capture important concepts.

There is no consistent approach or theoretical framework, however, in this broad-based research that one might use as a guideline to apply qualitative inquiry to the development of PROs [30–33]. Studies often provide frequency counts of very general themes, but focusing on frequency with such a small and varied number of subjects limits the informative value of qualitative research. Rarely (if ever) is a conceptual framework developed that could underpin an instrument. Clinical terms, such as "cancer-related fatigue," are often used to portray or define a concordance between the term and subjects' experiences. However, numerous studies exemplified by Schwartz and others have documented the discordant perception of many symptoms between subjects and their providers [34].

Overview of approaches in qualitative research

Qualitative research in the health arena has relied on several approaches to collect and analyze data, as well as interpret and present results. These mainly include phenomenology, ethnography, grounded theory, discourse analysis, and traditional content analysis. All approaches possess strengths and limitations, and they should be chosen dependent upon the type of research question(s) asked. Table 1 briefly describes the main distinctions between these major approaches; only phenomenology and grounded theory methods, the approaches we use, are more fully described in this article. Table 1 summarizes the essence, sampling methods, data collection and analysis method, and results yielded by commonly used qualitative methods in health science research. Please note that researchers use the names of these approaches interchangeably at times and have slightly different interpretations of them.

Using both an overarching phenomenological theoretical framework and grounded theory methods appears most suited to assuring the content validity and meaning of PRO concepts from the patient's perspective. In addition, it is most conducive to developing conceptual frameworks of questionnaires as required by the FDA. We chose these two approaches because they seemed the most appropriate to ensure the following:

- Using subjects' own words to describe their experiences rather than using broader themes developed through the eyes of the researcher to describe patients' experiences, as in traditional content analysis;
- Adequate sample size to ensure achieving saturation (no new concepts produced in final interview or sets of interviews); not generally a goal of narrowly defined phenomenology, ethnography, a case study, discourse analysis, or content analysis but important for PRO content validity and in grounded theory data collection and analysis methods; and
- The ability to develop items for a PRO measure rather than produce a narrative account of a subject's experiences within a social context.

Phenomenological theoretical framework

The scope for phenomenological research has been simply defined as "research designed to discover and understand the meaning of human life experiences" [35, p. 114]. Phenomenology seeks "to understand the lived experience of individuals and their intentions within their life world" [18, p. 24]. It answers the question, "What is it like to have a certain experience?" [18, p. 24] Although phenomenological inquiry underlies several qualitative approaches, the underlying belief is that the way to study a phenomenon is to access it through the eyes of the person(s) having the experience. This makes phenomenology the *sine qua non* overarching framework for PRO research. It is the theoretical underpinning that can guide research questions in a discussion guide and data collection, for example. Thus, a typical question is open-ended. It is neither constrained by preconceived theories nor taken from an instrument that may possess sound psychometric properties but was developed without subject input. In a concept elicitation interview, for example, on cancer-related fatigue one asks, "Please tell me about some of the symptoms you have experienced from cancer or cancer treatment?" rather than asking a question using the word fatigue, unless the subjects themselves have used it first. One probes for clarification on the meaning of responses like "tiredness," "weary," "weak," or "low energy." Participants are given ample time to express themselves.

Grounded theory methods approach

As the goal is an analysis that produces not only concepts but also a framework of items to be used as endpoints in clinical trials, the analysis and output of a pure phenomenological approach are insufficient. Using the umbrella of phenomenology, we suggest that grounded theory data collection and analysis methods best serve the development of a PRO structured questionnaire that can be used as an endpoint in a clinical trial.

Grounded theory is more a set of methods than a real theory. It can be seen as a "logically consistent set of data collection and analytic procedures aimed to develop

Table 1 Comparison of qualitative research approaches

	Phenomenology	Ethnography	Grounded Theory	Case Study	Discourse Analysis	Content Analysis
Essence	To understand the meaning of participants' experiences within their own "life world"	Immersion of researcher in setting to understand the ways of life of a cultural or social group	Set of data collection and analysis methods that assure that the meaning of a concept is discovered in the words and actions of participants from the ground up—not from application of *a priori* theory or concepts	To yield a full description or explanation of a phenomenon within a real life setting, e.g., an Alzheimer's unit	To describe how and why social interactions are routinely enacted using analysis of naturally occurring talk and texts (e.g., subject–physician interaction)	Researcher codes and abstracts into meaning units observational notes or transcripts of interviews, avoiding specific verbatim reports. Often uses prior theory frequency counts to describe prominent themes in text
Sampling	Few participants, usually $\leq$6, who have experienced the phenomenon	Key informants; observation of events; possibly participant observation	Progressive, as theory is built; number of participants depends on saturation; theoretical sampling	A case embedded in a single social setting but sampling of events, key actors, etc. occurs (purposeful sampling)	Random sampling of text, encounters, or sampling of social interactions	Observations or interviews
Data Collection	In-depth conversations in which interviewer brackets his/her own experiences from those of interviewee	In-depth and/or focus group interviews; observation	In-depth interviews with 20-30 participants, depends on homogeneity of participants; data collection continues until saturation achieved	Observations, archival data, interviews	Observation or recording of clinical interviews	Textual data from transcripts of interviews with participants, focus groups, or published documents
Data Analysis	Phenomenological reduction and structural synthesis; researcher identifies essence of phenomenon and clusters data around themes	Description, analysis, and interpretation of the social or cultural group; analysis may proceed in a number of ways including building taxonomies and making comparisons; often draws connections between the description of the group and broader extant theoretical frameworks.	Coding, sorting, and integrating data from verbatim report, and inductively building a conceptual framework to explain a phenomenon. Iterative process whereby further data collection is prompted by researcher's analytic interpretation; uses constant comparison method. Data collection stops when saturation of concepts achieved.	Reading through data- a transcript, notes, documents, objects; make margin notes and form initial codes; describe case and context; aggregate categories and discover patterns of categories; interprets and makes sense of findings	Transcripts analyzed with attention to minutia that might otherwise be considered "noise," e.g., hesitations, words such as "dunno," etc.; data are analyzed inductively and events and talk are seen as socially constructed through the interaction	Data usually coded into abstract codes and developed through the interpretative eyes of researcher; codes, concepts, or themes counted in terms of relative importance as seen by researcher
Results	Description of the phenomenon is often presented as narrative	Rich narrative description of cultural or social group, i.e., story with characters and a plot	A conceptual or theoretical model that describes concepts or categories and their relationships; usually presented as a visual graphic	Narrative augmented by tables, figures, and sketches	Description and explication of actions in everyday and institutional settings through analysis of talk or speech acts	Frequency counts of themes and descriptive quotes for a code

theory" [21, p. 27]. These qualities make it especially pertinent in the PRO field when capturing the dimensions of a concept and making transparent the development from verbatim concepts in textual data to item generation and development of meaningful domains [21, 36, 37]. It helps investigators conduct and analyze inquiry into a conceptual framework that can then be used to test quantitatively the reliability and validity of a PRO instrument.

Humanist assumptions underlie the use of grounded theory in the sense that it accounts for an index of feelings or meanings. Interviews are seen as representing an experience that allows access to authentic private selves, gives voice to the voiceless, and offsets the errors of positivism and prejudice [38]. Its founders, Nathan Glazer and Anselm Strauss, intended qualitative research to be a precursor to more rigorous quantitative research and wrote a clear set of guidelines [36].

According to these guidelines, there are three essential key characteristics of qualitative research: (1) it encourages participants to express their thoughts or feelings using very little structure during interviews; (2) it is iterative in the sense that concepts found in the data lead to other interviews to look for identical concepts or clarification of those concepts; and (3) its use facilitates development of a conceptual framework rather than substantiation of an *a priori* interpretation of a set of concepts.

Theoretical concepts used when conducting interviews emerge from the data. Charmaz's notions of changed self-identity of chronically ill subjects emerged from the data when, for example, a study participant with multiple sclerosis mentioned to her that when she was having a "bad" day, she dealt "with time differently and…that time had a different meaning to" her [39, p. 31]. This incorporation of concepts used by participants avoids the classic pitfall of taking for granted that the researcher shares the same meanings as the respondent [39]. Charmaz further explored the concept of good and bad days and found that good days indicated "minimal intrusiveness of illness, maximal control over mind, body and actions, and greater choices of activities"—all potentially important concepts in the experience and impact of chronic illness [39, p. 31].

Grounded theory methods are characterized by inductive (from the particular to the general) rather than deductive (from the general to the particular) reasoning. Data collection and analysis typically proceed in a simultaneous, iterative fashion. Researchers create analytic codes and categories from data, not preconceived hypotheses to which data might be overlaid. The researcher continues to use "memos," or thoughts or insights that they may have about the data during analysis. The term "memos" refers to analytic notes that the researcher writes during the coding process concerning the data, which serve as reminders of important thoughts and directions in which further data collection and analysis should go. Memos are conceptual and analytical rather than descriptive [37]. The researcher will search for codes, using a search function and Boolean operators (relationships defined by "and," "or," "not"), and develop models to explain the relationships between coded data. One could, for example, search for all quotes that contained the IBS codes "abdominal pain" and "cramping." All quotes that are given both of these codes can be output to examine the meaning of these pain concepts for patients: Is one really a sub-concept of the other? Are they two simple concepts? Does cramping describe severe abdominal pain? Is cramping in terms of pain descriptors, frequency, and location, the same abdominal pain? Are there any other pain sensations related to IBS? The researcher then compares and contrasts what different participants say to seek consensus (and deviant cases)—often referred to as the constant comparison method. When deviant cases are found, the researcher seeks to understand why. One may interview more subjects or re-analyze the data to search for clues that explain any deviations and their magnitudes. After comparing and contrasting the data multiple times (ideally with many researchers), the group would then develop a conceptual framework of concepts, associated sub-concepts, and items that might measure them. Examples of conceptual frameworks can be found in Patrick et al. (2007) and the Final Guidance to Industry [1, 6].

To make more clear how we applied our approach to a research question, we will use the study described below as an example in the rest of this article. In 2008, we conducted a focus group study to develop a PRO instrument for irritable bowel syndrome–constipation predominant (IBS-C) and irritable bowel syndrome–diarrhea predominant (IBS-D). There was a need for a new comprehensive measure of the signs and symptoms of IBS in which identified concepts would achieve saturation, concepts would be clear, the measure could capture clinically meaningful changes, and a meaningful responder definition could be defined. Our objective was to describe the experience of these conditions and their treatments and the impact on a subject's daily life.

Sampling strategies for PRO development

In a typical qualitative study, one uses theoretical or purposeful sampling rather than probability-based representative samples [21]. Representative experiences, not representative populations, are the goal. One may choose to interview several subgroups with different experiences of the condition. Depending on sample homogeneity, its size may be approximately ten to twelve persons per group. Relevant patient sociodemographic and/or clinical characteristics are used to sample and incorporate inclusion/exclusion criteria. These may include, for example, gender;

frequency, severity, and/or duration of symptoms; different treatment regimens; and treatments with different side effect profiles. Inclusion and exclusion criteria are very similar to the target population that is to be used in a clinical trial. Using the IBS example, one would try to ensure that the sample included mild, moderate, and severe cases, among other considerations, so as to understand the experience of having different levels of severity of IBS. This would allow the development of a PRO measure that could potentially capture both improvement and worsening.

Qualitative interviewing

Whether the researcher uses focus group methods or in-depth individual interviews, the development of a truly open-ended discussion guide is essential to ensure that qualitative interviews yield data that reflect the subject's experience without undue interviewer or discussion guide bias. The interview must be semi-structured, unlike the PRO items that eventually derive from it. The questions themselves do not have to be asked verbatim or exactly sequentially for every participant. Researchers might use sensitizing concepts from already-used instruments, other qualitative literature exploring the same phenomenon, acquired knowledge of the clinical condition, or interviews with clinicians. Some researchers advocate waiting until data collection and analyses are complete before conducting a literature review to avoid bias.

Four concept elicitation gender-specific focus groups were conducted and focused on understanding the symptoms of IBS-C that subjects experience. Gender-specific groups were chosen because of the sensitive nature of IBS. The interview guide was developed using sensitizing concepts found in the literature on IBS and in discussion with clinicians. Sensitizing concepts, a starting point for much of qualitative research, give guidance when approaching a phenomenon or experience but do not prescribe what the researcher should see; rather, they suggest where a researcher might want to look [40]. The moderator followed a semi-structured interview guide that included open-ended questions such as "What is a good day with IBS?" followed by "What is a bad day with IBS?" We were not interested in reporting what made a good day or bad day *per se* but in what symptoms and dimensions of those symptoms differentiated good from bad days.

Data analysis

Miller and Crabtree lumped grounded theory and phenomenology together in an "editing" data analysis style that incorporates the researcher as a text interpreter to identify units, develop categories, interpretively determine connections between such categories, and verify the initial data finding [18, p. 20]. This iterative process yields a final report that summarizes and details the data collection and analysis. Miller and Crabtree summarize this approach by metaphorically noting, "the interpreter enters the text much like an editor searching for meaningful segments, cutting, pasting, and rearranging until the reduced summary reveals the interpretive truth in the text" [18, p. 20]. While phenomenology provides a lens with which to view the experiences of subjects using narrative descriptions, grounded theory facilitates seeks to collect and codify experiences into a meaningful, conceptual framework. To develop meaningful PRO endpoints with content validity, researchers need to incorporate both approaches.

It is in the data analysis of qualitative research where grounded theory methods exhibit their strength for the development of conceptual frameworks. One builds these conceptual frameworks by induction, moving from specific to higher level concepts to even more general concepts (domains). In our research experiences, we have relied on Strauss and Corbin's [37] techniques and procedures. The researchers analyze as they collect data [37]. Here, the researcher engages in an iterative process, with the *modus operandi* coding system, and either manually codes "chunks" of transcribed text or "quotations," or does so using the increasing popular software packages that aid data analysis and theory building. Due to computerized data analysis, qualitative research has become more rigorous, efficient, and most importantly, transparent, while consuming less time.

In the IBS study, following each focus group, videotapes were transcribed verbatim. All identifying characteristics were removed from transcripts. First, the transcripts from the focus groups were entered into a qualitative software package, ATLAS.ti. ATLAS.ti is designed to facilitate the storage, coding, and retrieval of qualitative data using Boolean operators [41].

As the coding scheme is developed, some codes are repeated across interviews and some are not. For example, the first code a researcher may use when coding a transcript from an interview for IBS-D simply might be "pain." But upon examination of quotes that have been coded as "pain," different categories that represent a concept dimension may arise, e.g., "pain intensity," "pain frequency," "pain duration." As Charmaz summarizes, "As you raise the code to a category you begin: (1) to explicate its properties; (2) to specify conditions under which it arises, is maintained, and changes; (3) to describe its consequences; and (4) to show how this category relates to other categories" [21, p. 41].

Overall, the iterative and interpretive process of constant comparison analysis was used to develop or support a

conceptual framework for IBS-D and IBS-C. In this analytic process, subject quotations are compared and contrasted in several ways: iteratively by comparing earlier and later interviews; by sub-groups, for example, severity of the condition; and by concepts, e.g., whether "cramping" and "abdominal pain" are the same concept. Through such comparisons, it became clear that that abdominal discomfort, part of the clinical diagnostic criteria of both IBS-C and IBS-D, was a multidimensional concept. For subjects with IBS-C, discomfort meant a mild pain, fullness, and bloating. For subjects with IBS-D, abdominal discomfort appears to be an affective state (i.e., relates to an emotional response of feeling embarrassed) rather than a symptom itself that results from various symptoms and associated sensations (e.g., mild pain, bloating, and the immediate need to go). Subjects with IBS-D did not use the word "discomfort" per se but spoke of the uncomfortable aspects of IBS-D mentioned above.

A preliminary analysis on the transcripts was conducted to identify the concepts (i.e., root concepts or symptoms experienced by subjects with IBS) related to the research question. A list of every word and its frequency in the set of transcripts was generated. Each word was reviewed, and when a word appeared as a potential concept based on the team's knowledge of clinical indicators of IBS (e.g., abdominal pain), it was used to populate a list of root concepts. Word with same roots or which were conceptually equivalent were grouped together to shorten the list. This exercise started the coding scheme that reflected *potentially* important concepts based on the subjects' words, rather than being predefined by the researcher based on his/her knowledge of the condition, to avoid any bias [42].

Transcripts were assigned to different researchers for a thorough review preceding any coding to give the context of the subject's responses. Videos of the focus groups were observed by different researchers to look for nuances in body language and other visual cues. Visual cues, for example, included confirmation of the location of the body in which abdominal pain in IBS was experienced or the apparent anguish experienced with symptoms causing social embarrassment such as flatulence. A coding scheme with reliability was generated by the participation of more than one coder to process the first few transcripts, and subsequent group discussions of the interviews, transcripts, and codes. Code agreements and disagreements were discussed until consensus was achieved.

As researchers work with the text, they write memos in which they identify properties (characteristics of a category that defines or gives it meaning) and give their underlying assumptions about how categories develop or change either within a respondent's text or across respondents or time periods. When researchers coded subject responses to the good day/bad day questions, they would code, for example, for IBS-D severity: "…if you're going to have another bout of diarrhea later on in the day, or you're going to have stabbing pain while you're trying to do your job, and then have to leave"; or for frequency: "When you have diarrhea like three or four times." The researcher might write a memo stating,

It seems like subject 728 is talking about the feeling to trying to hold it. Lots of inaudible but it seems that 728, 723, 735, and 722 might be talking about the pain with the cramping. I need to check this assumption as I continue to collect and analyze data.

Such memos should be used to make comparisons initially between respondents, then categories, and finally concepts [36].

As they further compare and contrast data, researchers may need codes for subcategories to denote information such as the what, when, where, why, and how a concept is likely to occur [37]. This process will aid development of a conceptual framework that includes specific PRO concepts one wants to focus on in terms of potential treatment benefit. This interpretative process is the cornerstone of qualitative research, and it is necessary to condense the large volume of textual data. It is essential to capturing subjects' meaning of feelings and impacts of a condition and its treatment.

Theoretical saturation

According to Glazer and Strauss, researchers need to show that they have covered their topic in depth by having sufficient cases to explore and with which to elaborate their categories (or simple concepts) fully [36]. This is referred to as saturation. In the IBS study, the achievement of saturation was documented to show that all the concepts that were important for the subjects were considered for inclusion in the conceptual framework of a PRO instrument. Saturation is achieved if all concepts and their relationships with each other (how they may be grouped) are included in the conceptual framework. See Table 2 for a hypothetical framework for IBS-C.

The achievement of saturation ensures the adequacy of the sample size; if not achieved, new concepts emerge in the final focus group or interview, and further interviews must be conducted. When concepts and sub-concepts cannot be further specified with additional analysis or new data collection, saturation is achieved [37, 43]. The quantity of data in a category is not theoretically important to the process of saturation, and richness of data is derived from detailed description and not the number of times something is stated [43]. In qualitative research, "it is often the infrequent gem that puts other data into perspective that

Table 2 Example of conceptual framework for IBS-C

	Domains	Concepts
Primary symptoms	Constipation symptoms	→ Spontaneous incomplete bowel movement (SICBM)
		→ Complete spontaneous bowel movement (SCBM)
		→ Unsuccessful bowel movement (BM)
		→ Straining
		→ Stool consistency
	Other abdominal symptoms	→ Abdominal pain
		→ Bloating
		→ Abdominal fullness

Table 3 Hypothetical concept saturation of concepts in IBS-C focus groups

Concepts	Focus groups			Saturation
	1 vs. 2	1- 2 vs. 3	1-3 vs. 4	
BM consistency (liquid)	0 vs. 0	0 vs. 0	0 vs. 1	No
BM consistency (solid)	1 vs. 1	2 vs. 0	2 vs. 1	Yes
BM evacuation (incomplete)	1 vs. 1	2 vs. 1	3 vs. 1	Yes
BM evacuation (none)	1 vs. 1	2 vs. 1	3 vs. 0	Yes
BM frequency	1 vs. 1	2 vs. 1	3 vs. 1	Yes
BM size	1 vs. 1	2 vs. 0	2 vs. 0	Yes
Straining	1 vs. 1	2 vs. 1	3 vs. 1	Yes

Note BM color and BM odor were only mentioned by one subject, and mucus was mentioned by three subjects in only one focus group; therefore, these were not considered for inclusion in the saturation grid; rectal fullness and rectal pressure were related to incomplete bowel movement and straining, respectively

becomes the central key to understanding the data and for developing the model" [43, p. 148]. Table 3 shows an example of one way to present saturation.

Note that saturation is not a frequency count [36]. To graphically display the results of our evaluation of saturation we need to show that it is not a static but a dynamic concept. This graphic must display the iterative nature of qualitative data collection and analysis. To do so, the data were organized in chronological order, and the progression of concept identification within each focus group or interview was documented. Then concepts elicited across subjects were compared separately for each focus group using a stepwise approach: concepts elicited by the first set of subjects (focus group 1) were compared to the concepts elicited by the next set of subjects (focus group 2). The comprehensive list of concepts elicited from the first two sets of subjects was compared to concepts elicited from the third set of subjects (focus group 3); this process continued with the fourth focus group (see Table 3). A domain or simple clear concept was considered for saturation if the concept was elicited in at least one but not the last focus group or set of interviews and enough information was elicited to fully understand the meaning and importance of the concept to patients. If the concept was elicited only in the last focus group, then the saturation was considered questionable, and therefore, further data collection though focus group interviews would be recommended. The unit of analysis for the saturation grid was each focus group ($n = 4$) for the IBS study. For individual interviews the unit of analysis is preferably sets of interviews, for example 3 sets of five if 15 interviews were conducted.

Questions of reliability and validity of results

When qualitative researchers speak of validity, they are concerned primarily with credibility, transferability, and trustworthiness [44]. Sandelowski referred to validity as interpretive validity, where a "stable" category is confirmed by data [45]. Rigorous use of the procedures and techniques delineated herein, in conjunction with documentation of their use, will support the validity of the conceptual framework developed and the items that are formed from it.

The issue of reliability in qualitative research is controversial; however, working iteratively with teams to develop coding schemes and elaborating the data into categories, subcategories, and conceptual frameworks adds credibility to the notion that the results are reliable. In this sense, if another group were to collect and analyze these data in an identical manner, the outcome would be very similar to that in the initial study (reproducibility or repeatability). We suggest that one test the inter- and intra-rater reliability of the coding scheme as a measure of reliability. If resources and/or time prohibit this, one should have more than one coder process a transcript or use random samples of text from several transcripts to discuss any discrepancies pending consensus on a coding scheme. In the case of the IBS study, two senior researchers reviewed the coding and discussed any discrepancies between them. Kirk and Miller suggested that documenting the decision-making process of the research team as they work toward its conclusion allows the reader to evaluate the reliability of the results [46]. An example of a coding decision in the IBS study follows: Patients used the word "urgency" in both IBS-C and IBS-D and were coded with the same code. Further analysis suggested, however, that urgency of the immediate need to use the bathroom was a different concept in the two disorders. In IBS-D, the sense of urgency actually did mean the physical need to use the bathroom or was associated with fear of accidentally moving one's

Table 4 Key attributes of qualitative methods to develop PROs

Method	Key Attribute
Sampling	Representative of the experience
Interviewing	Open-ended elicitation of spontaneous responses
Analysis	Constant comparison; at least two coders; harmonization
Saturation	Iteratively achieved; not a frequency count; new concept does not add to conceptual framework
Reliability	Agreement between coders and within a coder's coding
Transparency	Documentation of the construction of the conceptual framework from the beginning of study

bowels. In IBS-C on the other hand, urgency was in effect feeling afraid of missing the opportunity to have a BM, a positive event for patients with IBS-C.

See Table 4 for a review of the key attributes of the qualitative methods presented in this article to develop PROs.

Triangulation

Triangulation refers to the combination of data sources, different researchers, multiple perspectives on a phenomenon of interest, or the use of multiple methods to arrive at conclusions about a research question [47, 48]. In qualitative research, triangulation gives greater perspective and allows for more credibility in one's findings. When the findings from methods and data sources converge, one has more confidence in them; when they diverge, this presents an opportunity to take a closer look at all data to gain a better understanding of the phenomenon being studied [47].

Findings from focus group data on IBS were triangulated (analyzed iteratively) with findings from cognitive interviews in IBS-C and another set of focus group data in IBS-D. Cognitive interviews consist of using verbal probing techniques to elicit respondents' thinking about items in a questionnaire to identify problems and support the content validity of questions [11]. In the IBS-C cognitive interviews, respondents' thinking regarding a set of IBS items was elicited, including their relevance, interpretation, and importance.

The inclusion/exclusion criteria, the demographic characteristics of the different study samples and the mean IBS severity level in each set of focus groups were compared. Finding the participants relatively similar in the different data sets, we continued with the triangulation process. Each data set was approached in the same way as described herein. A coding scheme was developed and harmonized. Coded quotations were compared and contrasted to develop concepts, sub-concepts, and domains. Saturation in both studies was evaluated, and the consistency of concepts and subjects' meaning between the datasets was confirmed.

Conclusion

This article sought to present methods to develop PROs through rigorous qualitative research. This not only fills a gap in the PRO literature but also moves beyond articles that suggest criteria to assess qualitative research in the health care field in general. Effective qualitative research is a crucial component of the objectives and requirements of PRO development and validation: to develop a conceptual framework; to use appropriate, meaningful, subject-friendly wording when developing the items, responses, and recall periods within questionnaires; and to test their face validity and comprehension; and finally, to ensure that no important items have been deleted (based on statistics alone) during the quantitative validation phase. Although researchers frequently omit this latter use of qualitative research or have overlooked its applicability, it is the crucial final step to ensure that the instrument possesses content validity. Rigorous, well-documented qualitative research provides evidence that the concepts, domains, and items in a PRO instrument are appropriate, comprehensive, and interpretable [49].

References

1. Patrick, D., Rock, E., O'Neill, R., Powers, J., Scott, J., Kennedy, D., et al. (2007). Patient reported outcomes to support medical product labelling claims: FDA perspective. *Value in Health, 10*, 125–137.
2. McColl, E. (2005). Developing Questionnaires. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials* (2nd ed., pp. 9–23). Oxford UK: Oxford University Press.
3. Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83–117.
4. Ware, J. E, Jr, Keller, S. D., Hatoum, H. T., & Kong, S. X. (1999). The SF-36 arthritis-specific health index (ASHI): I. Development and cross-validation of scoring algorithms. *Medical Care, 37*(5 Suppl), MS40–MS50.
5. Meyrick, J. (2007). What is good qualitative research? A first step towards a comprehensive approach to judging rigorous quality. *Journal of Health Psychology, 11*(5), 799–808.
6. US Department of Health and Human Services. (2009). Guidance for industry patient-reported outcome measures: Use in Medical Product Development to Suppor Labelling Claims.
7. Revicki, D. A., & Regulatory Issues and Patient-Reported Outcomes Task Force for the International Society for Quality of Life Research. (2007). FDA draft guidance and health-outcomes research. *Lancet, 369*(9561), 540–542.

8. US Department of Health and Human Services. (2006). Guidance for Industry: Patient-Reported Outcome Measures: use in Medical Product Development to Support Labeling Claims. Internet. http://www.fda.gov/cder/guidamce/5460dft.pdf.
9. Presser, S., Rothgeb, J. M., Martin, J., Singer, E., Lessler, J. T., & Martin, E. (2004). Methods for testing and evaluating survey questionnaires. *Public Opinion Quarterly, 68*, 109–130.
10. Tourangeau, R. (2000). Impact of cognitive models on survey measurement. In R. Tourangeau, L. J. Rips, & K. Rasinski (Eds.), *The psychology of survey response* (pp. 313–343). Cambridge, UK: Cambridge University Press.
11. Willis, G. (2005). Cognitive interviewing in practice. In Cognitive. Interviewing (Ed.), *A tool for improving questionnaire design* (pp. 42–62). Thousand Oaks, CA: Sage Publishers.
12. Kitzinger, J. (1995). Qualitative research. Introducing focus groups. *BMJ, 311*(7000), 299–302.
13. Pope, C., Ziebland, S., & Mays, N. (2000). Qualitative research in health care. Analysing qualitative data. *BMJ, 320*(7227), 114–116.
14. Hodges, B. D., Kuper, A., & Reeves, S. (2008). Discourse analysis. *BMJ, 337*, a879.
15. Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal of Quality in Health Care, 19*(6), 349–357.
16. Acquadro, C., Berzon, R., Dubois, D., Leidy, N. K., Marquis, P., Revicki, D., et al. (2003). Incorporating the patient's perspective into drug development and communication: An ad hoc task force report of the patient-reported outcomes (PRO) Harmonization group meeting at the food and drug administration, February 16, 2001. *Value in Health, 6*(5), 522–531.
17. Leidy, N. K. (2006). Evolving concepts in the measurement of treatment effects. *Proceedings of the American Thoracic Society, 3*(3), 212–217.
18. Crabtree, B. F., & Miller, W. L. (1992). *Doing qualitative research: Research methods for primary care*. London: Sage Publications.
19. Sandelowski, M. (2004). Using qualitative research. *Qualitative Health Research, 14*(10), 1366–1386.
20. Banister, P., Burman, E., Parker, I., Taylor, M., & Tindall, C. (2006). *Qualiative methods in psychology: A research guide*. Berkshire, UK: Open University Press.
21. Charmaz, K. (1995). Grounded theory. In J. A. Smith, R. Harre, & L. Van Langenhove (Eds.), *Rethinking methods in psychology* (pp. 27–49). London: Sage.
22. Shaw, R. (2001). Why use interpretative phenomenological analysis in health psychology? *Health Psychology Update, 10*(4), 48–52.
23. Smith, J. A., Harre, R., & Van Langenhove, L. (2005). *Rethinking methods in psychology*. London: SAGE Publications.
24. Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*(3), 229–238.
25. Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.
26. Schaeffer, N., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88.
27. Tourangeau, R. (2000). Respondents' understanding of survey questions. In R. Tourangeau, L. J. Rips, & K. Rasinski (Eds.), *The psychology of survey response* (pp. 23–61). Cambridge, UK: Cambridge University Press.
28. Tourangeau, R. (2000). Factual judgments and numerical estimates. In R. Tourangeau, L. J. Rips, & K. Rasinski (Eds.), *The psychology of survey response* (pp. 135–164). Cambridge, UK: Cambridge University Press.
29. Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). Washington, D.C.: National Academy Press.
30. Armes, J., Krishnasamy, M., & Higginson, I. (2004). *Fatigue in cancer*. Oxford, UK: Oxford University Press.
31. Dixon-Woods, M., Shaw, R. L., Agarwal, S., & Smith, J. A. (2004). The problem of appraising qualitative research. *Quality & Safety in Health Care, 13*(3), 223–225.
32. Wu, H. S., & McSweeney, M. (2004). Assessing fatigue in persons with cancer: An instrument development and testing study. *Cancer, 101*(7), 1685–1695.
33. Liamputtong, P., & Ezzy, D. (2005). *Qualitative research methods* (2nd ed. ed.). Melbourne, Australia: Oxford University Press.
34. Schwartz, C., Sprangers, M. A., & Fayers, P. (2005). Response shift: You know it's there, but how do you capture it? Challenges for the next phase of research. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials* (2nd ed., pp. 275–290). Oxford: Oxford University Press.
35. Girot, E. A. (1993). Assessment of competence in clinical practice: A phenomenological approach. *Journal of Advanced Nursing, 18*(1), 114–119.
36. Glaser, B. G., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.
37. Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed. ed.). London: Sage.
38. Sandelowski. (7-16-2000). 5th Annual Summer Institute in Qualitative Research, University of North Carolina School of Nursing, Chapel Hill, NC.
39. Charmaz, K. (1991). *Good days, bad days: The self in chronic illness and time*. New Brunswick, NJ: Rutgers University Press.
40. Bowen, G. A. (2006). Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods, 5*(3), 12–23.
41. Weitzman, E. A., & Miles, M. B. (1995). *Computer programs for qualitative data analysis*. London: Sage Publications.
42. Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. London: Sage Publications.
43. Morse, J. M. (1995). The significance of saturation. *Qualitative Health Research, 5*(2), 147–149.
44. Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report, 8*(4), 597–607.
45. Sandelowski, M. (2000). Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in Nursing and Health, 23*(3), 246–255.
46. Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. London: Sage.
47. Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). London: Sage Publications, Inc.
48. Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research*. London: Sage Publications, Inc.
49. Trentacosti, A. M. (2007). Epoetin Alpha: FDA Overview of Patient Reported Outcome (PRO) Claims. www.fda.gov/ohrms/dockets/ac/07/slides/2007-4315s1-09-FDA-Trentacosti.ppt. www.fda.gov/ohrms/dockets/ac/07/slides/2007-4315s1-09-FDA-Trentacosti.ppt.