Qualitative research and content validity: developing best practices based on science and experience

Meryl Brod · Laura E. Tesler · Torsten L. Christensen

Accepted: 13 September 2009 / Published online: 27 September 2009

Abstract

Purpose Establishing content validity for both new and existing patient-reported outcome (PRO) measures is central to a scientifically sound instrument development process. Methodological and logistical issues present a challenge in regard to determining the best practices for establishing content validity.

Methods This paper provides an overview of the current state of knowledge regarding qualitative research to establish content validity based on the scientific methodological literature and authors' experience.

Results Conceptual issues and frameworks for qualitative interview research, developing the interview discussion guide, reaching saturation, analysis of data, developing a theoretical model, item generation and cognitive debriefing are presented. Suggestions are offered for dealing with logistical issues regarding facilitator qualifications, ethics approval, sample recruitment, group logistics, taping and transcribing interviews, honoraria and documenting content validity.

Conclusions It is hoped this paper will stimulate further discussion regarding best practices for establishing content validity so that, as the PRO field moves forward, qualitative research can be evaluated for quality and acceptability according to scientifically established principles.

Keywords Assessing content validity · Patient-reported outcomes (PRO) development · Qualitative research

Establishing content validity for both new and existing patient-reported outcome (PRO) measures is central to a scientifically sound instrument development process. Content validity is "the extent to which one can generalize from a particular collection of items to all possible items in a broader domain of item … the intention is … to obtain as representative a collection of item material and relevant content as possible" [1]. Adequate assessment of content validity provides evidence that the conceptual framework, content of items and overall measurement approach are consistent with the perspective, experience and words of the patient group of interest and is necessary to meet FDA requirements for the development of PRO measures [2, 3]. Content validity is the measurement property that assesses whether items are comprehensive and adequately reflect the patient perspective for the population of interest. In addition, content validity provides evidence that formatting, instructions and response options are relevant, and the measure is understandable and acceptable to patients. Establishing content validity for PRO measures is critical as it supports the collection of appropriate and meaningful data, which can assist the health care system in providing optimal and relevant care to patients.

The most appropriate way to collect data to support content validity is by conducting qualitative research entailing direct communication with patients to adequately capture their perspective on issues of importance relative to the focus of the PRO measure. Both focus groups and individual interviews can be conducted to gather this information, with the collection and analysis of the information being systematic, documentable and qualitatively

M. Brod (✉) · L. E. Tesler
The Brod Group, 219 Julia Avenue, Mill Valley,
CA 94941, USA
e-mail: mbrod@thebrodgroup.net; meryl.brod@comcast.net

T. L. Christensen
Novo Nordisk A/S, Global Development, Krogshøjvej 29,
2880 Bagsværd, Denmark

accurate. Qualitative research covers a wide variety of conceptual principles and methodologies, incorporating the disciplines of sociology, anthropology, political science and psychology, and including wide variation in terms, concepts, assumptions and analytic principles. This variety of principles allows for tailoring of study designs to specific research purposes and can add great richness to data interpretation. Unfortunately, this variety is often viewed as a quagmire of confusing and sometimes contradictory study designs and methods, including case studies, politics and ethical inquiries, participatory inquiries, interviews, participant observation, visual methods and interpretive analysis [4]. Furthermore, qualitative data can come from three sources: interviews (individual and focus group), observations and documents. Qualitative data differ from quantitative research in that it considers the social and cultural construction of the variables of interest as integral to the concepts under objective examination, rather than seeking to correlate or factor out these influences. Therefore, it is more subject to bias than quantitative research and more difficult to structure. As a result, qualitative data have often been referred to as a "soft science" and the researcher as "a maker of quilts" [4]. Given the concerns about qualitative research being perceived as "soft science," [5–9] it is especially important to develop best practices that maintain the scientific integrity of the research process in order to maintain credibility. This rigor can be accomplished by having a sound scientific study methodology and protocol, including a semi-structured interview guide, appropriate analysis of the data and documentation of findings.

In addition to these conceptual difficulties, a variety of logistical considerations are crucial to the success of the research. Together, these methodological and logistical issues present a challenge to the PRO field in regard to determining the best practices of using qualitative research to establish content validity specifically in relationship to PRO development.

The purpose of this paper is to provide an overview of the conceptual and logistical "nuts and bolts" of conducting qualitative research to collect data to support the content validity of new and existing PRO measures, and present the published literature and theory on which these practices are based. Best practices for qualitative research must include both the conceptual and the logistical issues, as theory without implementation is not useful. To illustrate these issues, the development of a PRO to assess the impact of weight loss medication (WLM) for obesity will be used as an example. These best practices are based on the literature as well as the authors' experience, guided by the literature, conducting focus groups and individual interviews as medical outcomes researchers in both academia and industry. The first author is a PRO developer with a background in mental health and psychology, the second author is a medical anthropologist, and the third author has a background in economics and quantitative research implementation of PRO measures. We have tried to give practical and implementable information supported by examples based on experiences as well as guided by the theoretical qualitative research literature. It is our hope that this will serve as a springboard to further discussion of the best practices for the "how to" of using qualitative research to support content validity. The paper will focus on interviews, either conducted individually or with focus groups, as the data source to establish content validity for PRO development. However, observations and documents should not be ignored when considering patient groups who are not able to speak for themselves, such as the severely demented or very young. The paper is not intended to discuss interviewing techniques as they apply to ethnographic approaches to qualitative research.

These best practices are based on the grounded theory approach adapted for practical implementation in qualitative research and best suited for collecting data with high content validity for new and existing PRO measures. Grounded theory has been utilized in a number of studies evaluating various aspects of clinical trials, such as patients' experiences in and perspectives on participating in clinical trials [10–15]. Grounded theory has also been used to develop and/or evaluate outcome measures, for example, in dementia [16] and cerebral palsy [17]. In addition, it has been used to examine the roles of physicians and pharmacists in the pre- and postmarketing of new cardiovascular drugs [18].

The grounded theory approach supposes that theory is "grounded" in data, rather than presumed at the outset of the research [19]. In pure grounded theory, there would be no preconceptions of concepts of importance. Our approach adapts grounded theory, in that prior clinical knowledge based on expert opinion and the scientific literature is used as a starting point for domains and probes in the preliminary discussion guide. However, these probes are only asked after the unbiased first question of "Tell me about your experience with condition X," and domains and probes are changed based on the statements made by patients.

Grounded theory is based on two major principles: first, that phenomena are not conceived of as static but are rather constantly changing in response to evolving conditions; and second, that people have, although do not always use, the means of controlling their destinies by their response to conditions [20]. Although it is not the purpose of this paper to provide an in-depth and comprehensive review of qualitative research or grounded theory, there are some basic canons of grounded theory that should be understood, as they provide the scientific rationale for why grounded theory is especially relevant for PRO measure development and the framework for determining best practices for conducting the research. These tenets are

- Data collection and analysis are interrelated and concurrent, rather than linear processes; analysis begins as soon as the first bit of data is collected [21]. Accordingly, as emergent themes are identified in ongoing data analysis, all potentially relevant issues should be incorporated into the next set of interviews and observations [20, 21].
- Concepts are the basic units of analysis. Thus, data collected from subjects are given conceptual labels [20].
- Specificity of the concept is achieved by understanding the qualifiers of the concept (e.g., what factors impact the concept, such as age or gender).
- Concepts that pertain to the same phenomenon are grouped to form categories. Categories are further developed through repeated sampling (e.g., further interviews), examined in relation to one another and integrated into a theoretical framework [21].
- Analysis is achieved through constant comparison of similarities and differences in the data searching for both supportive and disconfirming evidence [20, 22]. Throughout the research process, hypotheses are revised based on the ongoing assessment of both qualifying and disqualifying evidence derived from interviews, observations and documents, until they can be fully supported by all of the data, facilitating a robust analysis. In addition, hypotheses must be repeatedly evidenced by the data, rather than based on a single instance [20]. Researchers must be aware of their own preconceived notions or biases in order to actively seek out data that challenges these [23].
- Sufficient data must be collected to reach "conceptual saturation," the complete elaboration of the properties, dimensions and variation that constitute each category or theme [24].

Applying grounded theory to PRO modeling, concepts are equivalent to PRO *subdomains,* and categories are equivalent to PRO *domains*. PRO domains, in turn, are grouped into an overarching *concept* (hence, this term constitutes the largest rather than the smallest unit of analysis in PRO development) [25]. In FDA terminology, each concept corresponds to a medical product labeling claim, which the PRO may be used to support [25].

Conceptual issues

Developing a new PRO is a multi-step process, and the techniques employed to obtain content validity, first in generating items based on the patient perspective and then for cognitive debriefing, are critical components of this process. Its place in the development process is shown in Fig. 1. The process of assessing and refining existing PRO measures will be discussed separately.

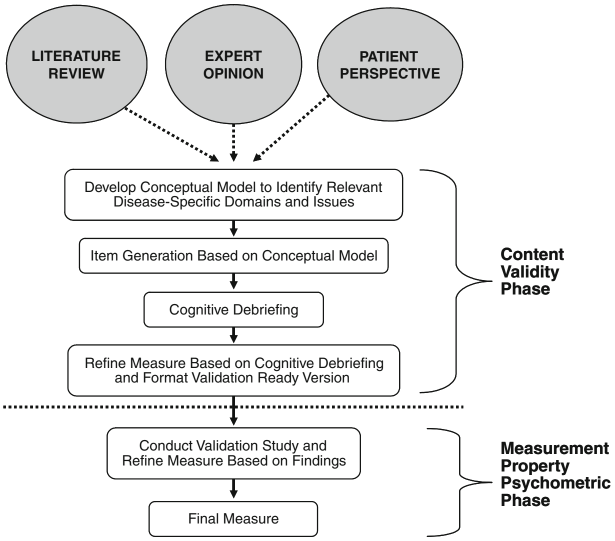
LITERATURE REVIEW
EXPERT OPINION
PATIENT PERSPECTIVE
Develop Conceptual Model to Identify Relevant Disease-Specific Domains and Issues
Item Generation Based on Conceptual Model
Cognitive Debriefing
Refine Measure Based on Cognitive Debriefing and Format Validation Ready Version
Content Validity Phase
Conduct Validation Study and Refine Measure Based on Findings
Final Measure
Measurement Property Psychometric Phase

Fig. 1 The PRO development process

The qualitative interview

The purpose of the interview process is to generate new information and confirm or deny known information. The process is an iterative one whereby each interview informs the next, and subsequent interviews are used to explore issues raised in previous interviews. The goal is not to reach consensus. An analogy to conducting qualitative interviews to assess content validity is the use of a GPS navigation system in cars. The GPS system can plot out your initial course, but it is also able to correct and revise when you make a wrong turn to get you to your final destination.

Both focus groups and individual interviews are valid and necessary techniques for collecting qualitative data. They should be viewed as complementary, rather than an either-or choice, as they may provide different information and allow the investigator to support the judgment that saturation has been reached independently of interview method. Focus groups help to identify "a range of experiences and perspectives," while individual interviews offer the opportunity for in-depth exploration [26]. Data may also be collected in less time and on a lower budget through focus groups compared with a series of individual interviews [27]. However, factors such as participants with mobility issues or other physical limitations, communication difficulties or time constraints may make focus group participation less feasible than individual interviews in some populations. There may also be practical constraints to assembling a group of individuals in one place at one time. Nevertheless, before limiting participation in a focus group due to such factors, thought should be given to conducting smaller focus groups, selecting venues more conducive to respondents with disabilities or conducting sessions of shorter duration [28, 29].

Focus group information is more influenced by group dynamic behavior and has the advantage of individual members having memories and thoughts stimulated or prompted by comments made by other members of the group [30]. It also provides a forum for participants to agree or disagree with each other, which facilitates the collection of disconfirming evidence. Analyzing how participants explain, defend or revise their perspectives to one another can provide insights into the factors that account for individuals' acceptance or rejection of certain ideas [27, 31]. However, focus groups also have the potential to "run away with themselves" and create a consensus of opinion, rather than idea generation. Some individuals may be reluctant to express their views or discuss certain experiences in a group, or they may defer to more dominant individuals and refrain from expressing a dissenting viewpoint. Focus groups may also be subject to polarizing effects, with individuals taking more extreme positions than they would in private [32].

Individual interviews provide a more private environment for patients to discuss and explore their own perspective without input from others. Individuals may also share more information about their experiences and perspectives in private interviews than they would in a focus group [31, 32]. In the absence of contributions from other group members, participants in individual interviews bear a greater responsibility for explaining and elaborating their statements, and frequently do so with little prompting [32]. Participants may also feel more comfortable discussing potentially sensitive or embarrassing subjects in an individual interview rather than in a group [31]. For some sensitive topics, however, facilitating a "segmented group" of individuals who share a key trait such as gender or the same disease may increase the comfort level of participants for the topic under discussion [33]. Thus, using a combination of both focus group and individual interviews provides a richer, more robust set of data to support content validity. In addition, conducting some individual interviews before, in between and after the focus groups allows the investigator to assess the generalizability of the information and obtain saturation across type of interview. Thus, it is an iterative and synergistic process between techniques.

Ideally, all interviews should be conducted by the same facilitator to help maintain consistency in elicitation and evaluation techniques across interviews. Further, having the same facilitator allows for more carryover from one interview to the next and for a more "organic" understanding of the issues. However, it is also beneficial to have an observer watch the groups, preferably from behind a screening glass so as not to interfere with group dynamics (group members should be informed that they are being observed). The observer can then assist the facilitator in later steps in the qualitative research process for data interpretation and analysis.

The interview guide

The structure and content of the interview should be based on a semi-structured interview guide or topic guide, which should be prepared before the first interview. The guide is developed from the researchers' prior knowledge of potential domains or areas of interest given the focus of the intended measure, literature review and expert opinion regarding the issues of interest. The questions in the guide are thus based on concepts or theory derived from either a priori information collected prior to the guide development or a working definition of the concepts of interest. For example, a guide looking at health-related quality of life issues would, at minimum, query respondents regarding the social, psychological, physical and symptom aspects of their condition [25, 34]. Although it is not relevant for respondents to understand the conceptual theoretical framework of the discussion guide, this information could be helpful to provide in the qualitative study protocol.

The guide is semi-structured in that it poses broad questions to the subject that can then be followed up through probes for further clarification. The flow of questions is funnel shaped, beginning with the most general to first gain an unbiased patient perspective, then to broad domains and finally to specific probes within a domain. The topic guide is the vehicle through which the researcher can achieve a balance between listening to the participant's story and questioning to elicit information about their experience with their health condition (e.g., symptoms, treatment satisfaction) under study as well as their social and psychological processes [21]. The semi-structured guide enables the facilitator to move around the guide using "emergent probing," and pursue avenues of discussion not in the guide or probe into topics that the subject brings up out of sequence from the guide [35–37]. Additionally, the guide can be, and is often, adapted between interviews as new themes or issues unfold from the interviews.

The facilitator must be flexible at all times to switch direction or topic from the guide while still covering all areas during the interview. Using the WLM PRO example, the first question might be "Tell me about your weight loss medication for obesity." Subsequent questions might be "Tell me about the social impact of taking obesity medication," with a probe question of "Does taking obesity medication impact your ability to go out with friends?" Additionally, questions such as, "Tell me what makes it easier/harder for you to take weight loss medication?" should be included. These types of questions help to identify factors (modifiers) that may either facilitate or hinder the relationship between WLM and outcomes. Questions that are geared to understanding the importance of a given impact should also be included so that the relative importance of an impact can be judged. Impacts that occur but are not meaningful to the subject should not be used to generate items for the PRO measure. Further, the facilitator should include queries that help identify the appropriate recall period and confirming attribution of symptoms or issues to the condition of interest. The discussion guide should end with a question, "Is there anything else I have forgotten to ask you about the impact of taking obesity medication," so that new information, not included in the discussion guide, can be considered for incorporation into the next interview.

Defining the sample

Theoretical sampling is a data collection strategy that relies on purposeful sampling of people, experiences and social phenomenon representative of and relevant to the topic of study, in order to identify and develop theoretical concepts (subdomains) and categories (domains). As new concepts emerge from the data, sampling criteria may evolve in order to more fully explore the range of variations, additional confirming or disconfirming cases, explanations for the social processes observed, and evidence of how concepts and categories relate to each other [22, 24].

Characteristics of the sample for focus group and individual interviews should reflect as closely as possible the patient population to be included in future studies that will incorporate the PRO [2]. However, within this range, as wide a distribution as possible of age, ethnicity and socioeconomic status is necessary in an effort to achieve a quasi-stratified, purposeful sampling where the sample is purposefully picked to represent a wide range of cases that demonstrate variation on both dimensions of interest and variations within a common group [22].

Samples may be either homogeneous or heterogeneous with regard to major patient characteristics. For focus groups, segmenting people into homogeneous groups helps to build a comparative dimension into a project (e.g., analyzing discussions by age, gender or class), and may facilitate the comfort level, sense of cohesiveness and/or flow of interactions among participants [26, 27, 32, 33]. Groups whose participants share similar characteristics tend to have greater compatibility, and dedicate more time to fulfilling the objectives of the session since less time is expended on group maintenance. They also experience less anxiety and greater satisfaction than those in incompatible groups [27]. A trade-off to conducting segmented groups, however, is that a greater number of groups must be conducted in order to reach a more representative sample [26, 32]. Homogenous groups may also limit the opportunity to explore differences in perspectives [30]. Research on single versus mixed gender groups, in turn, suggests that the latter may encourage greater participation, but due to the differences in the types of interactions that occur among participants, some researchers prefer to conduct focus groups of both types. Certain topics may also require more homogeneity since they are only relevant to certain segments of the population [27].

A focus group should be viewed as a temporary community of people with some similar characteristics (e.g., a common disease) who come together for a brief period of time to discuss that similarity. The differences in how that similarity is experienced and/or perceived become evident as the similarities are examined. The composition of the community is critical so that both similarities and differences become evident. Conducting the focus groups in different geographic locations to account for differences between urban and rural environments as well as regional variations in history, culture, health care access and practice issues will help to accomplish this. Additionally, variability in race, ethnicity, gender and age in respondents, as relevant for the condition under study, will facilitate identification of both the similarities and the differences in the community under study.

Of increasing importance is the applicability of PRO measures for international studies or for comparison of diseases across countries, thus requiring the community to include multiple countries so that the PRO developed is inclusive of all similarities and differences in the target population. As interviews should be conducted in the language of the country by a native speaker, it is generally not possible to have the same facilitator for each country. Identifying one facilitator as the "primary facilitator" who will train all other facilitators will allow for greater consistency among groups and improve data quality. Additionally, having the primary facilitator present at all groups, listening to the group by simultaneous translation and providing feedback to country-specific facilitators will further facilitate the "organic" understanding of issues under discussion by the primary facilitator and assist in the identification of similarities and differences between international communities.

Reaching saturation of new information

There are no power calculations or quantitative sample size estimations algorithms in qualitative research. Rather, to determine sample size, the investigator begins with a prespecified idea of a sample size based on the variability of the target population characteristics and clinical characteristics of the condition under study. Interviews should continue until "saturation" has been reached. This is the point whereby additional interviews are not expected to yield new or valuable information [38]. By this stage, sufficient data have been gathered to fully develop the depth and range of concepts (subdomains) and categories (domains) that explain the phenomenon under study, as well as an understanding of the relationships among concepts and categories [24]. Unfortunately, there are no clear cut rules on when "enough is enough," although it has been suggested in the literature that most projects reach saturation after conducting between 4 and 6 focus groups [26]. Previous research has found that after twelve interviews, between 88 and 92% of analysis codes (themes) can be identified [39]. One can determine the saturation point by making a qualitative judgment, supported by field notes as well as interview and focus group transcripts, that both key sample characteristics and concepts have been adequately sampled and that after conducting the prespecified number of interviews, no new information is being generated. If, after data analysis, it becomes evident that new important information emerged in the final group, then additional interviews should be conducted before a determination of reaching saturation can be made.

Preliminary judgments regarding reaching saturation can be made by the construction of a "saturation grid" whereby major domains (topics or themes) are listed along the vertical, and each group/interview is listed along the horizontal. This preliminary saturation grid can be constructed as the interviews proceed to help assist in the determination that saturation is likely to have (or not) been reached and make a determination as to whether additional groups will be necessary. Saturation is reached when the grid column for the current group is empty, suggesting that no new themes or concepts have emerged. However, the final determination that saturation has been reached is made during data analysis (coding) and documented with a refined saturation grid. In Table 1, the saturation grid includes the impact of WLM on the domains of Daily Life, Psychological and Treatment Burden. Each corresponding subdomain appears within the grid column corresponding to the focus group or interview in which it first emerged. Within the Psychological domain, the subdomains depression, anxiety and self-esteem were initially addressed by participants in the first focus group, while the subdomains anger and frustration were not addressed until the second group. In this example, saturation was reached for the Psychological and Treatment Burden domains, but not for the Daily Life domain. Therefore, at a minimum, another focus group would be required to reach saturation. This example represents our preferred technique for constructing a saturation grid. Alternatively, a saturation grid may be constructed by including all of the domains and subdomains along the vertical, and indicating with checkmarks or dots the sequences of coverage and points of emergence under each focus group or interview.

In our experience, a saturation grid based on field notes is highly correlated with the feeling of "I have heard all this before." Our rule of thumb, when combining both individual and focus group interviews, is that approximately 3–4 focus groups, in combination with 4–6 individual interviews, are generally sufficient to reach saturation whereby no new information is gained by further interviews. However, heterogeneity of sample, data quality, diffuse or vague areas of enquiry and facilitator skills will influence the exact number of interviews required to reach saturation.

The validity of the saturation grid is supported by quotes from the coded transcripts. In the Daily Life domain, the trouble focusing at work subdomain is supported by the quote, "You just can't stay focused, and usually the morning is my busy time when everybody's coming into work, and that's when that spike hits you. So, the worst time was the time that I would need it." In the Treatment Burden domain, the subdomain need to be near bathroom is supported by the quote, "I have to pick the places that I choose to go because I have to be near a bathroom" (Table 1).

Differences between establishing content validity for new measures versus existing measures

The major difference in establishing content validity for a new measure versus an existing measure is the goal of the interviews. For a new measure, the goal is to generate new information regarding the topic of interest based on previously identified possibilities, as well as newly provided information from the research participants. When assessing content validity for an existing measure, there are two goals: to determine whether the content of the existing measure is in fact relevant and important to the participants; and second, to assess whether there are additional areas of interest that are not covered in the existing measure. For an existing measure, it is possible to determine that the content validity of the measure is relevant and important, however, not inclusive of all information, thereby limiting the content validity and reducing the potential responsiveness of the measure.

Table 1 The saturation grid

Domain	Group 1	Group 2	Group 3
Daily Life	Hard to eat out with friends Don't want to be social when I can't eat Trouble focusing at work	Become annoying and people stay away Get more done when I am hyper	Hard to coordinate with work lunch meetings or business travel
Psychological	Depression Anxiety Poor self-esteem	Anger Frustration	
Treatment Burden	Inconvenient to remember to take Multiple pills a day are difficult	Need to coordinate with meals Need to be near bathroom Like pill form	

As with new measures, the discussion guide for qualitative interviews for existing measures should begin with a broad, open-ended question and then proceed to explore the theme of each of the items in the existing measure, rather than exploring domains and themes previously identified as potentially important. The third section of the interview guide should then explore additional areas of interest and/or query about other issues not captured in the measures items. It is also a good idea to have subjects complete the existing measure; however, this should be done at the end of the interview so as not to bias responses. The completed measures can then be examined to assess whether there are items that are either rated as not important or occurring rarely in the population of interest.

Analysis of data

Interviews should be transcribed verbatim, without editing to summarize or correct grammar and syntax, and should clearly indicate unintelligible speech. Attempts to decipher garbled speech should be enclosed in brackets with question marks [40]. Transcribers should use a notation system such as recommended by Poland [40] that clearly identifies pauses, emphasized words and expressions of emotion such as laughter and sighs. Transcribers should also indicate when a speaker is quoting or paraphrasing someone else. Conversational dynamics such as interruptions and overlapping speech should be preserved [40]. Examples from the WLM focus group transcripts are provided in Table 2. With focus groups, having a co-facilitator or observer present during the interview to take brief notes on who said what will help the transcriber to distinguish among the individual participants [41]. It is preferable to hire transcribers who have a social science and/or medical background and provide them with background information about the research scope and objectives to enhance their understanding of the subject matter. Interviewers and focus group moderators should review transcripts to check for accuracy and clarify ambiguities [40]. As with all members of the research team, transcribers should comply with Institutional Review Board and Research Ethic Committee protocols and maintain confidentiality to protect study participants.

The core tenet for the analysis of qualitative data is the ability to have pattern recognition achieved through content analysis of interview transcripts. Thus, it is during the analysis phase, rather than the interview process, that the consensus of issues should be reached. Inductive analysis involves discovering patterns, themes and categories that emerge, whereas in deductive analysis, the data are analyzed according to an existing framework [22]. Analysis to create a new PRO depends heavily on inductive reasoning, whereas analysis of existing measures relies more on deductive analysis.

Coding is the fundamental analytic process used to develop a theoretical conceptualization from the data. In grounded theory, there are three basic types of coding: open, axial and selective [20]. In open coding, the data are broken down into events, actions, interactions and emotions that can be compared for similarities and differences. Each of these is assigned a conceptual label so they can be grouped together into categories and subcategories. In axial coding, these categories are related to their subcategories. Finally, in selective coding, all categories are unified around an overarching core concept. In the development of a PRO instrument, open coding would result in the specific items of a measure; axial coding the domains; while selective coding would produce the overall concept the measure is intended to capture, thus closely reflecting the conceptual framework discussed in the draft FDA guidance [3].

Coding of data can be done either by "hand" or by computer software. Proponents of computer coding suggest that computer software programs are especially helpful for large amounts of qualitative data because they can reduce the amount of time that would otherwise be dedicated to

Table 2 Annotated excerpts from WLM focus group transcript

Denoted speech	Notation	Example
Paraphrasing others	Quotations are used to indicate that participant is mimicking what someone else has said.	F: And the doctor told me before he gave it to me: "You have to wear a diaper." You know, he was making a joke, like: "You have to wear a diaper with this. It'll work but it's messy."
Garbled speech (unable to make educated guess)	X's are used to denote unintelligible words. Each set of x's represent one word.	F: Oh, my goodness – he was right. It was really xxxxx xxxxx.
Garbled speech (able to make educated guess)	Brackets are used to denote unclear words for which the transcriber has guessed what was said, followed by a question mark.	Well, with the Phentermine, with the people who have anxiety and depression they give [Lexapro?], which is a depression and anxiety pill that's supposed to help that.
Interruptions	A double hyphen is used to indicate where someone's speech is interrupted by another.	I: Are any of you covered by health insurance for your - -? F: It doesn't cover it.

Notes: Notation style adapted from Poland [40]. In the transcript used for this table, participants were only identified by gender. F = female, I = Interviewer

manual and clerical functions and increase the thoroughness of handling data; unlike people, computers do not get tired or overlook passages of text. In addition, software programs can track changes and thus provide a more visible audit trail in data analysis. It has also been suggested that software programs can be used to identify complex relationships among coded concepts and categories, or other links in the data that might otherwise not be discernible [42].

However, computer programs are tools that *assist* analysis rather than analyze qualitative data. Qualitative software is above all a data management instrument, and its ability to actually code data is limited. Computer software programs can help the coder by facilitating data preparation, organization and management, but it is the researcher who identifies and defines the conceptual categories, determines the meaningfulness of the codes and interprets the theoretical significance of the data [42, 43]. Qualitative software facilitates the process but human beings do the analysis [20]. This is an important distinction, as reliance on computer analysis of data has a significant potential to produce erroneous results. The literature suggests that although software programs may save time, it is not clear if this adds to the analysis or detracts from it by distancing the researcher from the data through the mediation of computer, as the familiarity with the data engendered through repeated handling, reading and re-reading is an integral part of the analytical process [42, 44, 45]. Although more time consuming and a bit old fashioned, these authors still prefer hand coding as we feel it provides a more organic emersion and understanding of the data.

To code data:

- All statements in the transcripts are sorted into subdomains [45]. There are two types of subdomains—those preestablished by the questions of the semi-structured interview guide (e.g., sub categories of anger, embarrassment), and those that emerge from the data, which are not directly labeled by the interview guide. The subdomains that emerge as important (having an impact on overall concept and endorsed by a majority of subjects as important) serve as the basis for items in the measure. In the WLM example, subdomains that emerged from the data included frustration, depression, inconvenience to take medication and burden of frequency of administration.
- These subdomains are then grouped into categories, again either predetermined by the interview guide or newly emerging (e.g., anger and embarrassment categorized as psychological). These categories serve as the basis for the domains (which are operationalized as subscales of the measure). In the WLM example, based on the subdomains given above, the domains that emerged from the data were the burden of treatment and the psychological impact of treatment.
- The overarching concept captured by the domains is then named and is the basis for the final determination of the "type" or name of measure. This is what the overall measure "claims" to capture. In the WLM example, the overarching concept captured by the measure was labeled the treatment related impacts of WLM.

Both the individual domains and the overall concept may be suitable as the "claim," assuming that the appropriate psychometric testing is conducted to identify the factor structure, scoring algorithm and properties of reliability and validity.

It is recommended that someone other than the person who conducted the interviews independently do the initial coding of the transcripts during the analysis phase so that an independent final determination of reaching saturation can be made. After the initial coding, both the interviewer and the coder should meet to review the codes and reconcile any differences between them, reviewing coding themes and resorting the data as necessary. As with collecting data, the sorting process is iterative until both coder and facilitator are comfortable that they have appropriately captured the flavor and content of the interviews [46].

Developing a theoretical model and generating items

Based on the analysis, a theoretical model can then be developed. The theoretical model outlines the relationship between domains, consequences and modifiers. These modifiers can be considered as co-variates in statistical analyses for future studies that include the PRO data. The model helps to crystallize the factors that are domains of interest, this being important because only items that reflect domains should be included in the PRO. A preliminary model may be generated prior to the data analysis. However, the final model should be based on both the conceptual development and the psychometric validation findings once the preliminary domains have been confirmed by confirmatory factor analysis during the validation phase of the PRO development process. It should be noted that this theoretical model is distinct from the conceptual model term used by the FDA for the purpose of evaluating the validity of PRO-based product labeling claims. For the FDA, the purpose of a conceptual model is to identify and describe the specific PRO concepts and hypotheses that support each claim [47], whereas the theoretical model serves a broader purpose of helping to identify potential confounders to the relationship between the PRO measure and the other treatment outcomes in future studies which incorporate the PRO measure.

Using the WLM example, the theoretical model and the conceptual model would be as in Figs. 2 and 3.

Once the domains have been identified in the model, items reflecting the essence of the domain can then be generated as the first draft of the PRO. To support the assertion that items have high content validity, items generated should use the language of the subjects interviewed and directly reflect the content of qualitative statements made by subjects. For example, from the WLM example for the Side Effects domain, the quotes: “Jittery, insomnia biggest concern about staying on this … losing relationships and being avoided yes, feeling uncomfortable over time because of side effects;” or, “It feels like a hangover to me. It feels like a bad hangover. It feels like a hangover, like my head is just not all there and my stomach’s a little—I’m afraid to put certain things in it. I just don’t feel myself. I feel like I’m a shell, a jittery shell,” might result in the item: *Because of your weight loss medication, how often do you feel physical discomfort (for example, sleep problems, insomnia, jitteriness, diarrhea, gas, bloating, dry mouth)?* It is important when generating items to avoid bias in wording by using both anchors of a concept in the item (e.g., How easy or difficult is it to…) and at the same time wording questions so that the respondent attributes his or her response to the condition of interest (e.g., because of your…).

Using examples as a recall technique aid can be helpful when items may be ambiguous. Incorporating memory

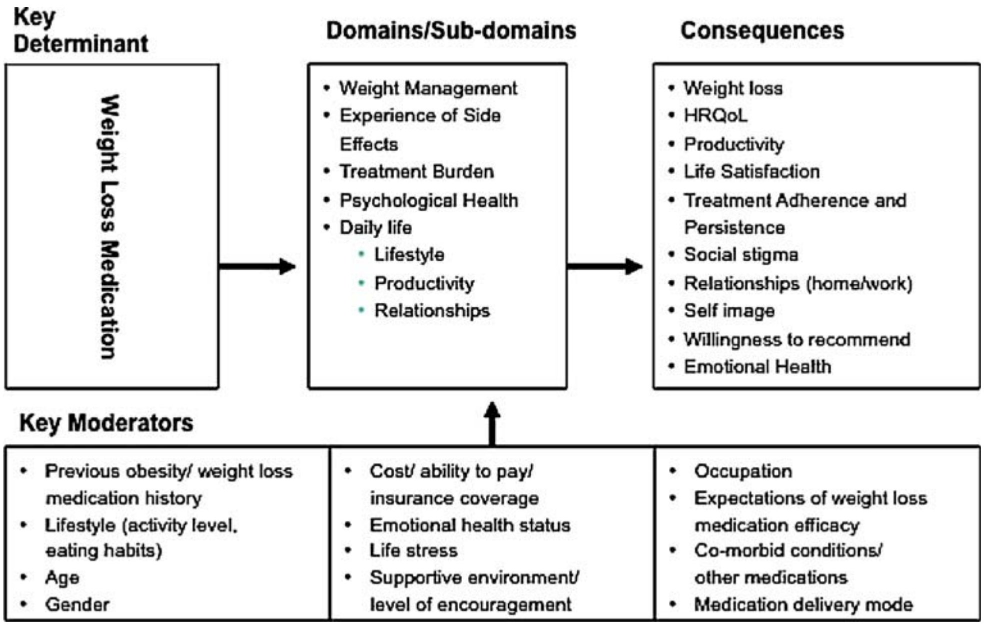
Key Determinant
Weight Loss Medication
Domains/Sub-domains
• Weight Management
• Experience of Side Effects
• Treatment Burden
• Psychological Health
• Daily life
• Lifestyle
• Productivity
• Relationships
Consequences
• Weight loss
• HRQoL
• Productivity
• Life Satisfaction
• Treatment Adherence and Persistence
• Social stigma
• Relationships (home/work)
• Self image
• Willingness to recommend
• Emotional Health
Key Moderators
• Previous obesity/ weight loss medication history
• Lifestyle (activity level, eating habits)
• Age
• Gender
• Cost/ ability to pay/ insurance coverage
• Emotional health status
• Life stress
• Supportive environment/ level of encouragement
• Occupation
• Expectations of weight loss medication efficacy
• Co-morbid conditions/ other medications
• Medication delivery mode

Fig. 2 Weight loss theoretical model

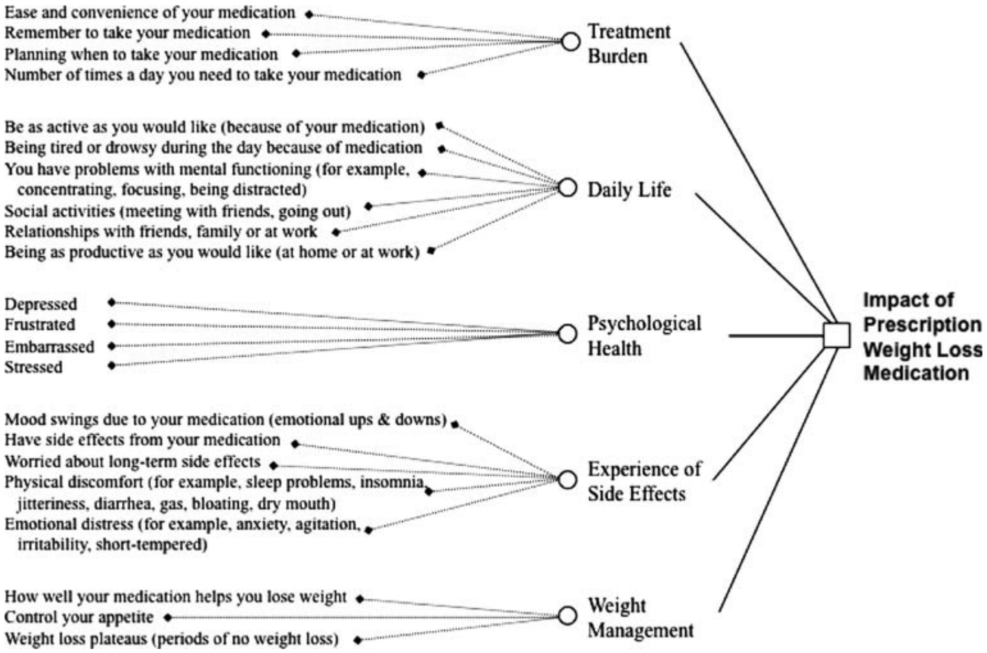
Ease and convenience of your medication
Remember to take your medication
Planning when to take your medication
Number of times a day you need to take your medication
Treatment Burden
Be as active as you would like (because of your medication)
Being tired or drowsy during the day because of medication
You have problems with mental functioning (for example, concentrating, focusing, being distracted)
Social activities (meeting with friends, going out)
Relationships with friends, family or at work
Being as productive as you would like (at home or at work)
Daily Life
Depressed
Frustrated
Embarrassed
Stressed
Psychological Health
Mood swings due to your medication (emotional ups & downs)
Have side effects from your medication
Worried about long-term side effects
Physical discomfort (for example, sleep problems, insomnia, jitteriness, diarrhea, gas, bloating, dry mouth)
Emotional distress (for example, anxiety, agitation, irritability, short-tempered)
Experience of Side Effects
How well your medication helps you lose weight
Control your appetite
Weight loss plateaus (periods of no weight loss)
Weight Management
Impact of Prescription Weight Loss Medication

Fig. 3 Weight loss conceptual model

cues into the question can also help participants to remember events that might not otherwise occur to them in the moment for events that might otherwise be forgotten [48]. For example, “How often does your medication interfere or not interfere with your social activities (meeting with friends, or going out)?” However, examples should be as inclusive as possible to avoid under-reporting relative to items that are mentioned specifically. If the number of alternatives is too great, the list may be restricted to a limited number of the most likely alternatives. Additionally, the list of examples should read “or” rather than “and,” which will reduce the likelihood of order effects or situations where one example may not be relevant and another not relevant to the respondent. For example, “Because of your weight loss medication, how often do you feel emotional distress (for example, anxiety, agitation, irritability or short-tempered)?”

Consideration of “translatability” of items into other languages should be kept in mind when generating items so that the intention of the item is consistent between translated versions. Tools for assessing likely translations of items are available [49].

The role of cognitive debriefing interviews in establishing content validity

Whereas the purpose of the interviews discussed earlier is to generate ideas, the role of cognitive debriefing interviews of the first draft of the measure is to reach consensus regarding the questionnaire format and structure to confirm that instructions and items are clear, understandable, inoffensive and relevant; that the recall period is appropriate; and that the format is acceptable to subjects [50]. Again, as with focus group and individual interviews, the sample for debriefing should match the characteristics of the intended study sample.

There are two primary approaches to cognitive debriefing interviewing. With the “think-aloud” technique, participants are asked to explain how they arrive at their answers to each question, during the actual administration of the questionnaire [51, 52]. Since the role of the interviewer is mainly limited to reading the survey questions, proponents of the think-aloud procedure believe that it minimizes the influence of the interviewer’s biases, standardizes the administration of the interview across participants and facilitates responses that provide new, unanticipated information. In addition, because the participant discusses his or her interpretation of each question during the actual survey administration, the responses may more accurately reflect the participant’s thought processes compared with post hoc interviewing relying on recall [51]. However, some participants experience difficulty in verbalizing their thought processes, or may go off on tangents, and the very act of talking their way through an answer may affect their response process [51].

In contrast with the think-aloud technique, an interviewer uses “verbal probing” after the participant has provided a response to a survey question. In this instance, the interviewer may ask participants to explain their interpretation of the question, how difficult or easy it was to answer the question and how they arrived at the answer. Follow up questions may also be asked to assess participants' recall time frames and vocabulary comprehension [52]. The structure and guidance provided by the interviewer through verbal probing may pose less of a burden on participants and help them to stay focused. Since reflection occurs after the participant has answered the survey question, proponents of this method argue that it prevents the cognitive debriefing process from affecting the answer. Additionally, the interviewer may elicit certain types of information or insights that the participant would not offer without being specifically asked to do so [51].

In sum, the think-aloud and verbal probing methods each offer a set of benefits and drawbacks. In the opinion of the authors, the think-aloud technique appears to be most useful for questions that require problem solving, while probing may be more appropriate for assessing a participant's familiarity with the subject matter and terminology posed in each question [35, 36]. Which technique to use may also depend on the amount of guidance required of each individual participant to discuss his or her thought processes and interpretation of the questionnaire, the subject matter of the questionnaire and whether it is to be self-administered [51]. Researchers should decide on a case-by-case basis the most appropriate method to use by clearly identifying the objectives of the proposed cognitive debriefing interviews and taking into consideration the content, format and mode of administration [35, 36]. In general, we have found that the verbal probing method is the most appropriate for cognitive debriefing interviews for PRO development.

Cognitive debriefing interviews are also an iterative process whereby interviews identify issues with the measure, items or instructions are reworded to eliminate the issue, and the revised measure is used for further cognitive debriefing interviews in a new sample. This process continues until consensus is reached that the measure is acceptable, resulting in a validation ready version of the measure [34]. Decisions to change an item, instructions or format are typically made when two or more participants have similar difficulty with some aspect of the survey task. Approximately seven to ten interviews are generally sufficient to reach consensus, depending upon the complexity of the instrument and the diversity of the participant population of interest [2].

Prior to conducting the cognitive debriefing interviews, an item definition list should be developed that states the intended meaning of each item so that the interviewer conducting the interviews can make a decision as to whether the subject understands the item as intended. The item definition table should list the item, the intended meaning and alternative acceptable meaning. This information will also be extremely valuable for translation efforts to help achieve content meaning equivalency between versions.

The steps in the cognitive debriefing interview process are as follows:

1. Subjects are contacted to obtain their agreement to participate.
2. Subjects are contacted at a prearranged time to conduct the telephone interview assessing readability, understanding and relevance of the measures. The time elapsed between steps 2 and 3 should be kept to an absolute minimum, preferably within 24 h.
3. After the first three subjects have been contacted, findings are reviewed, and a decision is made as to whether any changes to the PRO measure are required.
4. Three additional subjects are interviewed and edits made as required.
5. Steps one through five are repeated in blocks of three subjects each until there is consensus for a block of subjects that the measure is appropriate.

The authors have found that it is generally possible to email most interview respondents the PRO measure immediately before or at the beginning of the phone interview so that they can complete the questionnaire at the beginning of the interview. In cases where this is not possible, the PRO measure can be mailed in advance of the phone interview; however, respondents should be instructed not to complete the questionnaire until the phone interview, nor to discuss the measure with anyone else.

In general, the questions that should be asked in the cognitive debriefing interview include:

- In general, please tell me what you thought about the measure?
- What did the question mean to you?" [*compare respondent's definition to the Item Definition Table*]
- "Was the question worded in a way that made sense to you?"
- "Was the question in any way offensive or objectionable to you?"
- "Was the question about something which is important or relevant to you?"
- "Were the instructions and formatting clear?"
- "Did the response choices make sense?"
- "How did you select your response category"
- "Does the time frame you were asked to think about when answering the questions allow you to easily answer the question about your (condition)?"
- "When you completed the questionnaire, do you think you were able to accurately remember your experiences over the [*insert recall time frame*]?"
- "Is there anything we forgot to ask?"

- “Is there anything else you would like to comment on regarding the survey?”

After each question, the interviewer should then probe for the reason or explanation for the response and, if the response was a “no”, the interviewer should ask the respondent what alternatives they would prefer.

The formatting of the measure used for the cognitive debriefing interviews should be as close as possible, if not identical, to the format used in future trials. This will help reduce migration issues when going from a paper and pencil version to an electronically administered version.

Logistical issues

Facilitator qualifications

Given that there is an element of “art” in the science of qualitative interviews, the facilitator should be well trained in conducting qualitative interviews as well as in group dynamics. The consequences of a poorly trained facilitator can be disastrous, going beyond the collection of poor quality data, as often these subjects are ill, fragile and/or vulnerable. The facilitator has a responsibility to protect not only the confidentiality of the information, but also the emotional health of the subject. Thus, facilitators should be prepared and qualified to deal with a difficult, argumentative subject, emotional outpourings of fears and frustrations due to illness and/or questions regarding medical options. Additionally, the facilitator must be impartial, skilled and comfortable drawing out information from quiet members of the group, able to elicit a wide range of information without the need to reach consensus and able to maintain equal opportunity for contribution by all subjects involved. For example, a facilitator should be comfortable asking a subject to “please return to the topic under discussion” if they veer off course or ask a subject to “share the floor and give others a chance to speak” if they are dominating the conversation. Previous training in social work, therapy and psychology is helpful as these generally teach how to engage in active listening—a technique whereby the facilitator listens and then repeats and can probe without leading. Knowledge of the therapeutic area under study is also a benefit, although not a requirement, as it facilitates understanding of the discussion guide as well as prepares the facilitator for issues that may arise in the group that require “special handling” such as steering respondents away from irrelevant issues for the condition under study.

It is an added benefit to have an observer watching the group either from outside the room (most focus group facilities have observation rooms) or tucked away in the corner of the group room. In addition to being able to take notes to help the transcriber distinguish among individual participants, as discussed earlier, an observer fills the role of having another set of eyes and ears for the facilitator and can complete the saturation grid. This person also becomes extremely valuable in the analysis portion of the study, providing an additional perspective on the meaning and interpretation of the information collected. If the observer is watching from outside the room, the person’s presence must be made known to the participants.

The interview ground rules

It is important to begin the group with clear instructions as to the intention of the group and what the ground rules are. Basic rules should include the following: subjects are free to leave at any time; to please not speak when others are speaking; subjects should only speak about their own personal experience and not criticize or belittle experiences of others; no interrupting; speaking time is to be shared among participants and it is not the purpose of the group to give medical advice; participants should be referred to their own physicians for all medical questions. The ground rules should also make clear that the respondents as well as the facilitator should maintain confidentiality regarding statements made during the interviews.

Research Ethics Committee (REC)/Institutional Review Board (IRB) issues

Qualitative research should be subject to the same REC/IRB process as quantitative research; this includes obtaining REC/IRB approval. Ethics approval regulations as well as time frames for submission and approval differ between countries and should be addressed as country-specific issues. It is beyond the scope of this paper to discuss the various country-specific requirements. However, in general, qualitative research is generally viewed as low risk research. For example, in the US, given the nature of this type of study, an expedited review is often sufficient. Depending upon the subject recruitment and honoraria plan, written consent may not be required. Expedited review and waiver of written consent documentation can be requested from an IRB under FDA regulation 21 CFR 56.109(c), which allows for a waiver of written consent when the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside the research context. Focus groups are specifically listed as a minimal risk methodology (Category 7 of 21 CFR 56.110).

Requirements such as need for written consent will also vary by country. Using the US as an example, a consent letter not requiring signature is sufficient for qualitative interviews if a signed consent document would provide the only linking data between the subject and the data (45 CRF 46.117c). With a consent letter, consent is "implied," in that the subject reads the consent letter (or it is read before beginning a phone interview) and freely chooses to remain for the interview. A consent letter has all the same key elements of a signed consent form but does not require the principal investigator (PI) to be present when the letter is read (although the PI must be available for questions or concerns), nor the subject to sign the form. This greatly simplifies the IRB process in terms of documentation and time for approval. If a professional focus group/market research company is used for patient recruitment, a central IRB can be used to review and approve the project. It is wise to check with the IRB first to confirm that they are familiar with qualitative research submissions so that the process is not unnecessarily delayed due to misunderstandings about the study design's verbal or implied consent, as many IRB's have not reviewed these types of requests. Again, regulations will vary by country and should be investigated given the scope of a given study.

Recruiting the sample

A critical factor in recruiting the sample is to substantiate that the subjects have documentation that they have the condition/treatment of interest (matching the target population for future clinical trials). The easiest way to do this is recruitment of subjects through a physician. However, this is often costly and time consuming. First, a physician interested in participation must be identified. Next, the physician has to identify the eligible patients, contact them and obtain permission to give the interviewer their contact information. Finally, the interviewer must then contact the patient to arrange the interview/focus group. Alternatively, the physician may have the patient contact the interviewer. We have found that this approach will reduce the response rate as well as being time consuming. However, this may be the preferred option if identifying target population patients requires physician verification of clinician rated disease parameters such as severity. If physicians are used for recruitment, it is critical that no medical information regarding the patient be collected from the health care facility if expedited REC/IRB review without signed consent is expected, as collection of this information will invoke privacy protection regulations (e.g., HIPAA in the US). All medical information regarding the subject (e.g., medical history, medications taken) should be obtained directly from the patient and can be collected on a patient demographic and medical history form completed by the subject after the consent letter and before beginning the interview.

A more practical, less expensive and timelier approach is to use one of the many professional focus group/market research companies that recruit the subjects either from their database or from recruitment network, both in the US and internationally, although country-specific regulations and norms may apply and should be taken into consideration. Most facilities are able to recruit subjects within a 2-week timeframe. However, to substantiate that the patients are appropriate, it is necessary to have the facility require proof of condition/treatment from the subject. This can be in the form of a prescription, medication bottle, laboratory slip or letter from their physician as appropriate for the entry criteria. For example, for the WLM interviews, respondents would be required to bring to the interview either a recent prescription for WLM or a dated medication container. Additionally, if professional facilities are responsible for recruitment, the facility should be monitored for quality of the recruitment effort and special attention should be given to not recruiting people who are "professional focus group members," who are known to or report frequently attending focus groups.

Group logistics

Both day and evening groups, and at minimum evening groups, should be scheduled to accommodate working people. If the subject matter is delicate or of a sexual nature, it also may be best to conduct segmented male and female groups. The average length of time for a focus group is between 2 and 3 h, depending upon the sample (e.g., elderly patients may require a shorter group), and appropriate refreshments should be provided, especially for groups meeting around a meal time or in the evening, with people coming directly from work. Focus group facilities are generally prepared to provide a light meal. There is general consensus that 6–10 participants are the optimal size for a focus group. However, the actual size should depend upon issues such as the nature of the topic and characteristics of the participants (e.g., age and complexity of subject matter) where in such cases, smaller groups of 4–6 or 4–8 may be more appropriate [26–28, 30, 32, 53]. In our experience, the minimal number of subjects is generally three as below this size the focus group becomes cost prohibitive and interactions between participants become awkward. When fewer than three people can be recruited, individual interviews can be considered. Individual interviews can be conducted by telephone and, in the case of sensitive issues or when a highly widespread geographic distribution of subjects is desired, may be preferable to in-person interviews, which can be expensive and more difficult to schedule. Individual interviews can last up to 1 h, at which point subject fatigue begins to be a concern.

Recording and transcribing interviews

In order to document transparency and for analysis and review of data, all interviews should be recorded and transcribed. Options for recording include both audio and video, each with their advantages and disadvantages. Audio recording does not allow the transcriber to attribute comments to a given person (comments will only be identified as by male, female or facilitator) unless the facilitator names each person before they speak, a cumbersome and awkward process. Having the observer take brief notes on who is speaking or use of a real-time recording by a stenographer can assist in transcript interpretation and help to partially overcome this problem, although the added benefits given the budget implications should be considered. Video eliminates this problem, but it is more likely that subjects may refuse to be recorded. Additionally, video recording raises more complicated REC/IRB issues as subject identification then becomes possible via the recording. Given these issues, we generally prefer audio recording. Although it can be beneficial to attribute comments to a specific individual, in the authors' opinion, it is generally sufficient to identify participants by gender only. Focus group facilities are well prepared to tape groups and several options exist for professional recording of telephone interviews via phone companies. As discussed in REC/IRB issues, regardless of interview type, the subjects should be informed prior to the interview that they are being recorded and offered the opportunity to refuse. It has been our experience that this rarely occurs. In less than ideal recording environments (outside of professional facilities) a back up digital recorder and/or double recordings will provide a safety backup for recording mishaps.

Recordings of interviews should be transcribed for use in the data analysis. It is helpful to have transcripts include mention of nonverbal behaviors such as laughter or crying; however, more subtle paraverbal interpretations such as disgust or anger can generally be identified by words and are, therefore, less likely to be inaccurately transcribed. Mentions to other group members should be restricted to first names only to maintain confidentiality. The transcript should be reviewed for accuracy by comparing the text with the interview tape or by careful review by the facilitator or interview observer shortly after the group to avoid recall bias. Professional transcribers can generally produce a transcript of a 2 h group within 24–48 h. Recordings should be securely stored and, the authors suggest, saved for the same period of time required of all clinical trial data in accordance with country and institution-specific regulations and guidelines. In the US, the guidelines require that clinical trial records be retained for 2 years after the latter of the following dates: the date a marketing application has been approved for a medical device or drug for the specific indication under investigation, or the date an investigation has been discontinued and the FDA notified [54, 55]. The transcript itself should also be considered confidential and handled accordingly.

Honoraria

It is customary to give subjects an honorarium for completing an interview. This honorarium should be included in the REC/IRB submission request and must be commensurate with the level of effort of the subject so as not to be considered coercive. At the time of writing this paper, an acceptable honoraria in the United States is approximately $125.00–$175.00. Additionally, subjects may be reimbursed for any travel expenses to attend an interview. Some REC/IRB's may request that the honoraria be slightly less for the telephone interview than for focus groups, as less effort is required on the part of the subject in terms of time and travel.

Documenting content validity

It is essential to the task of evaluating content validity that the process be transparent and well documented for both scientific and regulatory purposes. Intuitively, establishing content validity is part of the measure development as it is the process of identifying patient-relevant issues and generating items that reflect those issues. However, assessing content validity is a psychometric property and in reality is not truly confirmed until the measure is used in a study where there is also a successful intervention and the measure is shown to be responsive to that change.

There is general consensus in the scientific PRO community that the methodological standards for development of a PRO are similar across countries. However, regulatory differences between FDA and EMEA do exist and should be considered depending upon country-specific regulatory needs. The FDA, for example, recommends that documentation for PRO instruments used in clinical trials include the protocol for qualitative interviews and focus groups, cognitive debriefing interviews and any other research used to identify concepts, generate items, or revise an existing instrument, including training of interviewers, the qualitative interview strategy, description of qualitative interviews and focus groups, transcripts, coding procedures and justification for each version of the developing instrument to support its adequacy [25]. In the current version of draft guidance for the FDA, there are several sections where content validity information is relevant, including Sects. IV.A.1, Identification of Concepts and Domains; IV.B.1, Generation of Items; IV.B 3, Choice of Recall Period;

IV.B.5, Evaluation of Patient Understanding; and IV.C.2, Evaluation of Validity. Our rule of thumb, until further guidance is given, has been to put text regarding content validity in Sects. IVA and B as appropriate to the section and then in Sect. IVC2 refer the reader back to previous sections. In the text, we suggest including the sample description and all information regarding sample selection, methods and analysis strategy and results. The table documenting changes made during cognitive debriefing, the final validation ready version of the measure (postcognitive debriefing) and the last iteration of the semi-structured interview guide can be placed in the appendices. Transcripts of the interviews, analysis coding of themes, and definition of items used for the cognitive debriefing are offered as available upon request. Other published papers that have addressed the issue of documenting content validity are available [25, 47, 50, 56].

Peer-reviewed manuscripts can provide another avenue for documentation of the research. The manuscript should follow the same basic outline of any scientific article including introduction, methods, results and discussion. Manuscripts that outline the patient-reported conceptual issues surrounding illness and medication can provide valuable insight to health care providers for the condition of interest and make a significant contribution to clinical understanding.

Conclusions

Qualitative research to establish and support content validity should have a strong and documentable scientific basis and be conducted with the rigor required of all robust research. This rigor is supported by an interviewer who is well versed in conducting scientific qualitative research and who understands the importance of accurately reflecting the patient voice. Meaningful qualitative research requires art as well as science. Accurate collection of patient data, as well as analysis of that data, requires a level of empathy and understanding of patient issues along with the ability of the researcher to “listen and interpret,” which is not required in quantitative research and can often not be easily taught. As a result, conducting qualitative research is not to be undertaken lightly lest we risk the development of PROs with poor responsiveness and clinical meaning. The task for establishing and assessing content validity is to lay out a set of best practices so that at minimum, PRO qualitative research can be evaluated for quality and acceptability. This paper is intended to present both the scientific evidence and our experience regarding current thinking on best practices. We welcome future debate on these practices.

References

1. Nunally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed., p. 104). McGraw-Hill: New York.
2. Leidy, N., & Vernon, M. (2008). Perspectives on patient-reported outcomes. Content validity and qualitative research in a changing clinical trial environment. *Pharmacoeconomics, 26*(5), 363–370.
3. U.S. Department of Health and Human Services. (2008). Food and drug administration. Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims. Rockville, MD. http://www.fda.gov/cder/guidance/index.htm.
4. Denzin, N., & Lincoln, Y. (Eds.). (2003). *Collection and interpreting qualitative materials* (2nd ed.). Thousand Oaks, CA: Sage.
5. Snape, D., & Spencer, L. (2004). The foundations of qualitative research. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 1–23). London: SAGE.
6. Theobald, S., & Nhlema-Simwaka, B. (2008). The research, policy and practice interface: Reflections on using applied social research to promote equity in health in Malawi. *Social Science and Medicine, 67*, 760–770.
7. Friedland, G. H. (2006). HIV medication adherence: The intersection of biomedical, biobehavioral, and social science research and clinical practice. *Journal of Acquired Immune Deficiency Syndromes, 43*(Suppl 1), 53–59.
8. Greenhalgh, T., & Taylor, R. (1997). How to read a paper: Papers that go beyond numbers (qualitative research). *British Medical Journal, 315*, 740–743.
9. Firestone, W. A., & Herriott, R. E. (1983). The formalization of qualitative research: An adaptation of “soft science” to the policy world. *Evaluation Review, 7*, 437–466.
10. Belue, R., Taylor-Richardson, K. D., Lin, J., Rivera, A. T., & Grandison, D. (2006). African Americans and participation in clinical trials: Differences in beliefs and attitudes by gender. *Contemporary Clinical Trials, 27*, 498–505.
11. Featherstone, K., & Donavan, J. L. (1998). Random allocation or allocation at random? Patients’ perspectives of participation in a randomised controlled trial. *BMJ, 317*, 1177–1180.
12. Lawton, J., Fox, A., Fox, C., & Kinmonth, A. L. (2003). Participating in the United Kingdom prospective diabetes study (UKPDS): A qualitative study of patients’ experiences. *British Journal of General Practice, 53*, 394–398.
13. Madsen, S. M., Holm, S., & Riis, P. (2009). Attitudes towards clinical research among cancer trial participants and non-participants: An interview study using a grounded theory approach. *Journal of Medical Ethics, 33*, 234–240.
14. Marsden, J., & Bradburn, J. (2004). Patient and clinician collaboration in the design of a national randomized breast cancer trial. *Health Expectations, 7*, 6–17.
15. Paterniti, D. A., Chen, M. S., Chiechi, C., Beckett, L. A., Horan, N., Turrell, C., et al. (2005). Asian Americans and cancer clinical trials: A mixed-methods approach to understanding awareness and experience. *Cancer Supplement, 104*(12), 3015–3024.
16. Silberfeld, M., Rueda, S., Krahn, M., & Naglie, G. (2002). Content validity for dementia of three generic preference based health related quality of life instruments. *Quality of Life Research, 11*, 71–79.
17. Waters, E., Maher, E., Salmon, L., Reddihough, D., & Boyd, R. (2005). Development of a condition-specific measure of quality of life for children with cerebral palsy: Empirical thematic data reported by parents and children. *Child: Care, Health and Development, 31*(2), 127–135.

18. Wieringa, N. F., Peschar, J. L., Denig, P., de Graeff, P. A., & Vos, R. (2003). Connecting pre-marketing clinical research and medical practice. *International Journal of Technology Assessment in Health Care, 19*(1), 202–219.
19. Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine Press.
20. Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons and evaluative criteria. *Qualitative Sociology, 13*(1), 3–21.
21. Charmaz, K. (2003). Qualitative interviewing and grounded theory analysis. In J. A. Holstein & J. F. Gubrium (Eds.), *Inside interviewing: New lenses, new concerns* (pp. 311–330). Thousand Oaks, CA: Sage.
22. Patton, M. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
23. McGhee, G., Marland, G. R., & Atkinson, J. (2007). Grounded theory research: Literature reviewing anf reflexivity. *Journal of Advanced Nursing, 60*(3), 334–342.
24. Corbin, J., & Strauss, A. (2007). *Basics of qualitative research* (3rd ed.). Newbury Park, CA: Sage.
25. Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P., Dawisha, S., et al. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health, 10*(Suppl 2), S125–S137.
26. Morgan, D. (1996). Focus groups. *Annual Review of Sociology, 22*, 129–152.
27. Stewart, D., Shamdasani, P. N., & Rook, D. W. (2006). *Focus groups* (2nd ed.). Thousand Oaks, CA: Sage.
28. Quine, S., & Cameron, I. (1995). The use of focus groups with the disabled elderly. *Qualitative Health Research, 5*(4), 454–462.
29. Koppelman, N., & Bourjolly, J. (2001). Conducting focus groups with women with severe psychiatric disabilities: A methodological overview. *Psychiatric Rehabilitation Journal, 25*(2), 142–151.
30. Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *BMJ, 311*, 299–302.
31. Greenbaum, T. (2000). *Moderating focus groups: A practical guide for group facilitation*. Thousand Oaks, CA: Sage.
32. Morgan, D. (1997). *Focus groups as qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.
33. Hollander, J. (2004). The social contexts of focus groups. *Journal of Contemporary Ethnography, 33*(5), 602–637.
34. Turner, R. R., Quittner, A. L., Parasuraman, B. M., Kallich, J. D., Cleeland, C. S., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). Patient-reported outcomes: Instrument development and selection issues. *Value Health, 10*(Suppl 2), S86–S93.
35. Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage.
36. Willis, G. B. (2004). Cognitive interviewing revisited: A useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 23–44). New York: Wiley-IEEE.
37. Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 45–66). New York: Wiley-IEEE.
38. Cutliffe, J. (2000). Methodological issues in grounded theory. *Journal of Advanced Nursing, 31*(6), 1476–1484.
39. Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59–82.
40. Poland, B. (2003). Transcription quality. In J. A. Holstein & J. F. Gubrium (Eds.), *Inside interviewing: New lenses, new concerns* (pp. 267–288). Thousand Oaks, CA: Sage.
41. Bernard, H. R. (2005). *Research methods in anthropology* (4th ed.). Walnut Creek, CA: Rowman Altamira.
42. St John, W., & Johnson, P. (2000). The pros and cons of data analysis software for qualitative research. *Journal of Nursing Scholarship, 32*(4), 393–397.
43. Jennings, B. (2007). Qualitative analysis: A case of software or 'peopleware?'. *Research in Nursing and Health, 30*, 483–484.
44. Morison, M., & Moir, J. (1998). The role of computer software in the analysis of qualitative data: Efficient clerk, research assistant or Trojan horse? *Journal of Advanced Nursing, 28*(1), 106–116.
45. Ritchie, J., Spencer, L., & O'Connor, W. (2003). Carrying out qualitative analysis. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 219–262). London: Sage.
46. Hruschka, D., Schwartz, D., St John, D., Picone-Decaro, E., Jenkins, R., & Carey, J. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods* 307–331.
47. Rothman, M. L., Beltran, P., Cappelleri, J. C., Lipscomb, J., Teschendorf, B., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). Patient-reported outcomes: Conceptual issues. *Value Health, 10*(Suppl 2), S66–S75.
48. Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions*. New York: John Wiley and Sons.
49. Acquadro, C., Conway, C., Wolf, B., Anfray, C., Hareendran, A., Mear, I., et al. (2008). Development of a standardized classification system for the translations of patient-reported outcome (PRO) measures. *Quality of Life Newsletter, 39*, 5.
50. Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., & Hays, R. D. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health, 10*(2), S94–S105.
51. Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 1–25.
52. Willis, G. B. (1999). Cognitive interviewing: A "how to" guide. Resource document. National Cancer Institute. http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf. Accessed 2 May 2009.
53. Krueger, R. (1995). The future of focus groups. *Qualitative Health Research, 5*(4), 524–530.
54. U.S. Department of Health and Human Services. (2008). Food and drug administration. CFR—Code of Federal Regulations Title 21: Part 812—Investigational Device Exemptions. Resource Document. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?fr=812.140.
55. U.S. Department of Health and Human Services. (2008). Food and drug administration. CFR—Code of Federal Regulations Title 21: Part 812—Investigational New Drug Application. Resource Document. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?fr=312.62.
56. Revicki, D. A., Gnanasakthy, A., & Weinfurt, K. (2007). Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: The PRO evidence dossier. *Quality of Life Research, 16*, 717–723.