Qual Life Res (2012) 21:739–746
DOI 10.1007/s11136-011-9990-8

REVIEW

Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting

Susan Magasi · Gery Ryan · Dennis Revicki ·
William Lenderking · Ron D. Hays · Meryl Brod ·
Claire Snyder · Maarten Boers · David Cella

Accepted: 3 August 2011 / Published online: 25 August 2011

Abstract Content validity of patient-reported outcome measures (PROs) has been a focus of debate since the 2006 publication of the U.S. FDA Draft Guidance for Industry in Patient Reported Outcome Measurement. Under the auspices of the Patient Reported Outcomes Measurement Information System (PROMIS) initiative, a working meeting on content validity was convened with leading PRO measurement experts. Platform presentations and participant discussion highlighted key issues in the content validity debate, including inconsistency in the definition and evaluation of content validity, the need for empirical research to support methodological approaches to the evaluation of content validity, and concerns that continual re-evaluation of content validity slows the pace of science and leads to the proliferation of study-specific PROs. We advocate an approach to the evaluation of content validity, which includes meticulously documented qualitative and advanced quantitative methods. To advance the science of content validity in PROs, we recommend (1) development of a consensus definition of content validity; (2) development of content validity guidelines that delineate the role of qualitative and quantitative methods and the integration of multiple perspectives; (3) empirical evaluation of generalizability of content validity across applications; and (4) use of generic measures as the foundation for PROs assessment.

Keywords PRO development · Content validity · Qualitative research · Quantitative research

Abbreviations

CAT	Computer adaptive testing
CFA	Confirmatory factor analysis
DIF	Differential item functioning
FDA	United States Food and Drug Administration
IRT	Item response theory
ISOQOL	International Society for Quality of Life Research
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
NIH	National Institutes of Health
PRO	Patient-reported outcome
PROMIS	Patient Reported Outcomes Measurement Information System
SEM	Structural equation modeling

S. Magasi (✉) · D. Cella
Department of Medical Social Sciences, Feinberg School
of Medicine Northwestern University, 625 Michigan Ave.,
Suite 2700, Chicago, IL Il 60611, USA
e-mail: s-magasi@northwestern.edu

G. Ryan
Rand Corporation, Santa Monica, CA, USA

D. Revicki · W. Lenderking
United BioSource Corporation, Bethesda, MD, USA

R. D. Hays
Department of Medicine, University of California, Los Angeles,
Los Angeles, CA, USA

M. Brod
The Brod Group, Mill Valley, CA, USA

C. Snyder
Division of General Internal Medicine, Johns Hopkins School
of Medicine, Baltimore, MD, USA

M. Boers
Department of Epidemiology and Biostatistics,
Vrije Universiteit (VU) University Medical Center,
Amsterdam, The Netherlands

Introduction

The *AERA/APA/NCME Standards for Educational and Psychological Testing* [1] assert that validity is a unitary concept and "the degree to which evidence and theory support the interpretation of test scores entailed by the proposed uses of the tests" (p. 9). As such, AERA/APA/NCME no longer advocate subdivisions of this unitary concept according to traditional nomenclatures (such as content, construct or predictive validity). There has, however, been increasing emphasis on the traditional concept of content validity in the field of patient-reported outcome (PRO) measures, particularly when PRO measures are recommended to support regulatory labeling claims of patient benefit from new treatments. Interest in content validity was heightened by the 2006 publication of the U. S. FDA Draft Guidance for Industry on Patient-Reported Outcome Measures [2]. Published 3 years later as a final guidance [3], this document asserted the primacy of the patient in evaluating content validity and demands the solicitation of direct patient input to ensure that appropriate end point measures are used to support pharmaceutical and device labeling claims. According to the FDA guidance, patient input during instrument development and item selection will help ensure that PRO end points are directly relevant to the life and disease experiences of the target population [3].

The Patient Reported Outcomes Measurement Information System (PROMIS) initiative has sought to advance measurement using a combination of interview-based qualitative methods and advanced psychometric methods including item response theory, item banking, and computerized adaptive testing (CAT) [4]. PROMIS investigators sought to measure core self-reported health concepts like pain, fatigue, depression, anxiety, sleep, physical function and social function across health conditions, age, and gender. Specific manifestations of these concepts may characterize given diseases or treatments and can be evaluated clinically and empirically. Therefore, according to PROMIS, just as there is not a disease-specific measure of blood pressure or hemoglobin, there may not be disease-specific fatigue or pain. This reflects a *domain-specific* rather than *disease-specific* approach to measurement. PROMIS' generic domain-specific approach to assessing symptoms and functioning has, however, raised questions about content validity and generalizability of health concepts when measured across conditions, populations, and settings.

Under the auspices of PROMIS, a 1½ day working meeting on content validity in PRO measures was convened in Chicago, IL, USA in July 2009. The purpose of the meeting was to develop definitions and practice guidelines to advance the science of content validity in PRO measures. Participants included leading experts on PRO measures from both inside and outside the PROMIS network, including representatives from the U.S. Food and Drug Administration (FDA), National Institutes of Health (NIH), International Society for Quality of Life Research (ISOQOL), International Society for Pharmacoeconomics and Outcomes Research (ISPOR), and patient advocates (see attached participants list). The meeting consisted of platform presentations and extensive participant discussion and debate. This article is divided into two sections. The first half provides an overview of the platform presentations and outlines key conceptual and methodological issues related to content validity as presented in the meeting by the co-authors. The second half closes with suggestions for evaluating content validity, including recommendations for future research. These recommendations reflect the summary of the meeting presentations, including rather extensive discussion. They are not, however, presented as formal consensus statements, as this was not the intent of the meeting. While both the Chicago meeting and this article emphasize content validity, the authors recognize that evidence based on a PRO measure's content is only one piece of the accumulated evidence needed to evaluate a PRO measure's validity. It is not the intent of this manuscript to elevate content validity above other sources of validity evidence but rather to acknowledge and respond to the current regulatory environment and state of the science.

Defining content validity

To understand content validity, it is helpful to examine fundamental aspects of PRO measure development, including (1) specifying the *concept* the scale is intended to represent; (2) *scaling* the concept's various components and items; and (3) defining the PRO measure's intended *purpose*.

Concepts

PRO measures are designed to measure underlying, often invisible health concepts such as pain, fatigue, depression, and physical function. PRO measures are more than a collection of items; they must be conceptually related to an underlying and evidence-based conceptual framework. A concept's boundaries need to be determined and can be subjected to discussion and interpretation. Concept boundaries are often defined qualitatively by triangulating patient perspectives, expert opinion, and literature review. Once the concept has been defined and bounded it must be operationalized through the development of individual items, a process typically done by instrument developers and informed by content expertise, legacy measures, and qualitative data. During the item development process, it is important to remember that "nearly all tests leave out elements that some potential users believe should be measured and include some elements that some potential users consider inappropriate" [1, p. 10]. Theory and empirical evidence must be considered together to maximize instrument validity.

Eliciting a concept's components and potential items

It has been argued that qualitative research is the most appropriate way to achieve and assess content validity as it allows direct communication with patients [5, 6]. Qualitative methods allow the instrument developers to capture patient perspectives on the concept of interest and evaluate both comprehension and acceptability of PRO items. Interviews (either individual or focus group) are the preferred method for elucidating the patients' experiences as they might be captured in a PRO measure.

Grounded theory is perhaps the most appropriate approach for content validity research [5]. Grounded theory is an approach to "conducting qualitative research that focuses on creating conceptual frameworks grounded in the data" [7, p. 187]. A basic tenet of grounded theory is that data collection and analysis are interrelated and iterative processes. Therefore, analysis occurs throughout the data collection process, and research findings should inform and enrich subsequent interviews and observations. Data analysis within a grounded theory approach involves the identification of themes and conceptual categories through constant comparisons for similarities and differences in the data across individuals, groups, and contexts. Nuanced understandings of the concept and its operationalization into conceptual frameworks and items are thus achieved by understanding both the core concept and its qualifiers [8, 9]. It has been argued that disease-specific measures or items are qualifiers of more general/generic measures [5].

Qualitative data collection should be based on a purposive theoretical sampling strategy founded on theoretically or conceptually based demographic and clinical variables that are hypothesized to influence the disease experience. A well-targeted sampling strategy helps ensure "efficient and effective saturation of categories, with optimal quality data and minimum dross" [10, p. 12]. To ensure that qualitative findings are representative, researchers should ensure that the concept is adequately described (and reflected in the items developed) by bringing participants into the study until no new themes are obtained [11]. This point of redundancy is known as theoretical saturation. The number of interviews required to reach saturation depends on the complexity of the health concept being described. It has been argued that saturation may be achieved in as few as 6–12 interviews [12] but for practical purposes most studies set the sample size at 20–30 interviews [13]. Researchers typically do not embark on the instrument development process naïve to the concepts they are seeking to measure. They bring clinical, conceptual, or content expertise to the process. This expertise, coupled with a thorough review of the literature, informs the sampling strategy, data collection and analysis [8, 11].

The relative weight of patient versus expert input is central to the content validity debate. For example, even when patient input is used to develop item content, someone must decide what remains and what is removed from a questionnaire. Anthropologists have long struggled with this problem and have characterized a distinction between *emic* and *etic* viewpoints [14]. *Emic* explanations are based on insiders' perspectives and understandings of how things work. *Etic* explanations are based on outsiders' perspectives and understandings. Although emic and etic approaches may produce distinct explanations, both have their purpose and they are often complementary.

When applied to content validity, the concepts of emic and etic stimulate several important questions. For instance, who determines which components should be considered part of a concept? Who decides which items are selected as part of the final scale? Who evaluates whether the scale is appropriate for a given application? These questions force us to make explicit the roles of insiders versus outsiders. Insiders—such as patients—often have the advantage of first-hand experience of the concept. Outsiders—such as clinicians, researchers, or subject matter experts—often have the advantage of having observed how the concept manifests across a wider array of people and contexts and have often systematically collected data on the similarities and differences across such people and contexts. Optimal content validity does not generally emerge when one perspective is prioritized over the other.

The FDA and ISPOR guidance documents [3, 13] emphasize the role of the emic (patient) perspective in the PRO development and selection process. Yet, decisions regarding content validity are often made from the outside (etic) perspective of test developers or review agencies/panels. While PRO developers could take a completely emic or etic approach to data collection, in practice most developers include some combination of both. It is important to make explicit the sources from which items were derived and how they were prioritized for the instrument development process. Data collection and analysis procedures should be documented and transparent to potential end users and regulators [5, 13].

Scaling

Once a concept's components have been identified and items developed, the items should be field tested, and the empirical data can be analyzed quantitatively to identify unidimensional or multidimensional scales. Scaling a concept involves selecting items to align them along one or more continua and estimating the scale's precision. Which items are dropped and which are retained, and how these actions shape a concept's boundaries, carries implications for the content validity of a scale. Concept under representation and concept-irrelevant variance are the two threats to validity that the patient's voice is designed to address. Contemporary measurement models and item response theory (IRT) posit that the items measuring a health concept are hierarchically arranged from easiest to hardest. Once items are calibrated along this hierarchy, it is no longer necessary for all patients to answer all questions or even the same set of questions. Rather by capitalizing on CAT technology, items may be targeted to pinpoint a patient's experience of a health concept with a small subset of items. Item calibrations also provide a mechanism for linking PRO measures that use different items and response scales to yield comparable scores. Situational and patient factors, such as developmental stage, may influence how health concepts are expressed, yet the underlying concept remains comparable.

While qualitative methods help define the boundaries of a concept and the initial item content, quantitative analyses provide insight into content validity of multi-item and multi-dimensional PRO measures. Quantitative methods can be used to explore and confirm the dimensionality and structure of multi-item scales; evaluate item bias across demographic groups; and examine relationships among health concepts. Results of quantitative psychometric analyses confirm and extend the qualitative research findings. To ensure an instrument's validity, developers would benefit from avoiding a false dichotomization of qualitative and quantitative methods and adopting iterative mixed methods approaches. Confirmation of content validity is dependent on the accumulation of research evidence. However, once sufficient evidence from multiple sources is demonstrated, it is reasonable to conclude that there is enough information on content validity of the targeted PRO. Thus, quantitative evidence is a critical part of the iterative process in developing content valid PRO measures.

In the PRO instrument development process, exploratory and confirmatory factor analysis can evaluate items for fit within a hypothesized domain by demonstrating that items with a specified domain scale load onto the factors [15]. Items that cross-load on multiple factors may be removed from an instrument. Factor analysis allows us to understand the internal structure of a PRO measure and to evaluate the consistency of the factor structure across different samples. Factor analysis also leads to the development of summary scores from multi-domain measures. Confirmatory factor analysis (CFA) within the context of content validity allows for the evaluation of specific hypotheses about factor structures and content and can be used to hierarchically test for invariance in factor structure across groups [16, 17].

Structural equation modeling (SEM) provides the framework from which CFA can be used to evaluate the extent to which concepts being measured across groups of people have the same characteristics and boundaries in relation to other concepts. Once measurement equivalence is established, SEM can be used to evaluate structural (e.g., group) equivalence. SEM is also used to confirm hypothesized factor structures of PRO measures [18, 19]. SEM is used to evaluate the factorial validity of PRO measures by confirming specified relationships between the PROs and antecedents and consequences of interest. SEM allows researchers to assess multiple domains simultaneously and examine the longitudinal relationships between clinical and PRO end points. Finally, SEM can be used to cross-validate PRO measures across subgroups (i.e., gender, language versions, etc.). SEM allows for the evaluation of complex relationships between clinical and PROs [20]. For example, these models have been used to examine the longitudinal relationships between treatment-related impact on hemoglobin in patients with chemotherapy-induced anemia and the effect of changes in hemoglobin on changes in patient-reported fatigue [21]. SEMs can also be used to evaluate and confirm PRO end point models.

Measurement invariance can also be evaluated using differential item functioning (DIF). DIF examines the relationship among item responses, levels of a concept being measured, and subgroup membership [22, 23]. For any given level of a concept, the probability of endorsing a specific item response should be independent of group membership. There are two kinds of DIF. Uniform DIF is consistent across the range of the concept being measured, and non-uniform DIF varies depending on the concept level. DIF testing can be done with ordinal logistic regression. DIF is identified as a significant effect of subgroup membership on item score after controlling for the level of the concept. The concept level is approximated by summing across items or estimating IRT scores.

IRT information curves show the investigator where an item bank or instrument is not covering the continuum of severity or impairment [22]. This information is useful in targeting when additional development work and domain content coverage is needed.

Linking key components and items pre- and post-scaling

During the psychometric scaling process, some items are usually eliminated. Here, we need to ask "How (if at all) has the original concept been narrowed or changed as the result of such eliminations?" In this phase of the scaling process, there is an inherent tension between creating a scale with valid psychometric properties and being inclusive in terms of the content covered. How this tension is resolved is partially dependent on how the scale will be used. The most reasonable way to address content validity at this stage is to make explicit how the content might have narrowed. Describing which (if any) key components were dropped and noting if those dropped had been elicited primarily from insiders, outsiders, or both gives end users a better understanding of what the scale represents and what it does not. Additionally, qualitative methods such as member checking provide a means of ensuring that quantitatively determined decisions, such as the removal of misfitting items from a calibrated item bank or distillation of an item bank to a short form, still adequately capture the participants' experiences.

Purpose

Throughout the concept definition and scaling processes, one must bear in mind the purpose and context for which the scale will be applied. It is important to consider if and how the application of PRO measures to novel contexts and populations impact validity claims. For example, can a generic measure of pain be valid across different disease and diagnostic groups? Other examples of changing contexts include use of the PRO measure in different clinical populations, cultural settings or countries, or clinical practice versus research applications.

In sum, *content validity is the extent to which a scale represents the most relevant and important aspects of a concept in the context of a given measurement application.*

Ultimately, documenting content validity entails making explicit the process by which items are selected relative to the concept they are supposed to represent. This involves describing from where they came and who was involved in both identifying and ultimately selecting them. Only when such information is made available are researchers and end users in a position to adequately judge whether a scale's content is appropriately aligned with both the concept it is supposed to measure and the use to which it will be put. Generalizability, or the ability to apply measures developed for one clinical population to another, emerges as an important issue when users of PRO measures seek to use validated PRO across clinical contexts and populations. It is not, however, feasible to test every possible combination of demographic and disease-related factors qualitatively for every subpopulation in which a measure might be used.

Recommendations and conclusion

Establishing the validity of a PRO measure for a particular purpose (such as the measurement of depressed mood as a method for evaluating the efficacy of an antidepressant) is an ongoing process, requiring both qualitative and quantitative approaches. It is not possible to arrive at the level of 100% certainty, but it is possible to estimate the probability of having a valid instrument. Perhaps, the cause of some of the issues around the interpretation and application of the FDA guidance to instrument development and interpretation of clinical trial results comes from the fact that the FDA operates in a regulatory context, which requires application of a deterministic, legally bound code to the decision-making process regarding fitness for purpose. In contrast, the science of human measurement that is being regulated is based on probability and approximation. In any case, elevating content validity to pre-eminent status in the process of evaluating the validity of instruments is (a) not consistent with the well-accepted standards in the field of psychometrics and (b) runs the risk of evaluating instruments based on subjective opinions (in this case of regulators), which emphasize the voice of the patient and which more than 100 years of quantitative methodological development were designed to avoid. The acceptability of a PRO measure should not depend upon satisfaction of every single standard in a literal manner, but should rest upon professional judgment, the degree to which standards have been met, the availability of alternatives, and the feasibility of satisfying the standards. There are multiple methods to evaluate and ultimately establish the content validity of PRO measures. These include qualitative methods such as the use of patient focus groups, individual patient interviews, and expert input; and quantitative methods to evaluate the generalizability of measured concepts across patient populations. A fundamental unanswered question pertains to the extent to which content validity findings are generalizable across clinical samples, social and environmental contexts. When it comes to defining the state-of-the-science of content validity, there are clearly more questions than answers. Our focus was to bring together experts on the topic to develop a short list of recommendations to advance the field. For these recommendations, we draw on decades of research in health and outcomes measurement [1, 24]. The AERA/APA/NCME Standards for Educational and Psychological Testing, in particular, has elaborated detailed criteria for document validity evidence [1].

Recommendations

1. We recommend the *adoption of a consensus definition of content validity*. We propose the following definition as a starting point: "Content validity is the extent to which a scale or questionnaire represents the most relevant and important aspects of a concept in the context of a given measurement application". A

consensus definition of content validity will help organize and direct the research agenda. This definition is broad enough to include various opinions on the topic; and yet, it includes some key unanswered questions that would benefit from further research.

2. We recommend the *development of content validity guidelines* that acknowledge the role that both emic and etic perspectives play in the PRO development process.

   Although direct patient input helps to ensure the meaningfulness of conceptual models and PRO measures, it is impossible to come to closure and produce a good PRO measure without incorporating expert (etic) perspectives. The current regulatory environment's emphasis on patient perspectives fails to appreciate the influence that expert opinion and psychometric analysis have on PRO measure development and content validity. Explicit articulation of the role that different stakeholders play in the instrument development process would enhance future users' abilities to evaluate validity claims. Furthermore, content validity guidelines must include "decision rules" on how to reconcile differences when patient and expert opinions diverge. Adopting an integrated mixed methods approach to instrument development and evaluation process will help assure the PRO measures are both psychometrically robust and content valid for the given measurement application.

3. We recommend that *generalizability be assessed by empirical research*. Generalizability of a PRO measure to patient groups not previously studied is a critical issue in the content validity debate. We are concerned about a growing implication that even rigorously developed and evaluated measures need to be "validated" each time they are used in new clinical groups. An alternative perspective seems equally reasonable: Well-developed measures of common symptoms or functional problems may be valid across various groups. This perspective can justify a questionnaire as "reasonable for use" across patient groups, while at the same time testing the assumption of generalizability. When checking content validity across clinical groups, questions will emerge, such as "How does one determine the optimal threshold for what is a "relevant" component?" and "at what point do subgroup differences regarding the relative importance of a component matter at the level of measurement?" These and other questions cannot easily be addressed without further basic research comparing item development, item selection, and scoring methods. For example, are core concepts experienced differently across populations? How about a disease syndrome across different demographic groups or a symptom across different diseases (e.g., is fatigue in diabetes different from fatigue in AIDS, and if so how)? While these are important empirical questions, we contend that perhaps the most salient question is not documenting if a PRO is content valid for each potential group of end users but rather a clear examination of what renders a PRO content *invalid*? What clinical and/or conceptual basis exists to challenge content validity once it has been documented through the rigorous application of qualitative and quantitative methods?

4. We recommend the *use of generic measures as the foundation for PRO assessment* in clinical trials, with disease-targeted measures used to supplement as needed.

   By defining content validity in relativist terms, we risk perpetuating the tendency of researchers to develop study-specific outcomes measures because of a perceived lack of "validated" measures that directly match their target sample. The need to continually re-certify a measures' content validity is not only onerous for the research community, it is bad for science and evidence-based practice. It can lead to a proliferation of measures all purporting to measure the same underlying concept but with potentially divergent items, thereby limiting the ability to compare findings and outcomes across settings, samples, and populations. For these and other reasons, we recommend that well-developed and validated generic measures can be regarded as "fit for purpose" in specific clinical settings where content validity has not been previously documented.

   A well-calibrated item bank within a CAT application can contain within it a broad enough range of items to capture the symptom experiences and impacts for a variety of clinical populations. Disease-targeted symptoms may be seen as qualifiers of generic measures. Within the context of item banking and CAT applications, the issue of scalability was deemed highly relevant. Namely, can we assume that targeted short forms derived from content valid item banks retain the validity across settings? The group advocated for empirical research to evaluate validity claims across clinical populations. For example, empirical study of fatigue experience and impact in diverse clinical groups such as people with cancer, people with MS, and people with chronic fatigue syndrome. The inherent value in the application of generic measures of fatigue, such as those developed as part of PROMIS is evident for such comparative research.

Conclusion

Rigorous application of mixed methods research, including meticulously documented qualitative methods and advanced quantitative analyses, provides PRO developers and end users with tools needed to evaluate and document content validity. Some important empirical questions remain. For example, when and why do content validity claims need to be re-evaluated across different patient populations? Put another way: When can we conclude that accumulated knowledge and data about a PRO measure is sufficient to support an assumption of content validity regarding the measured concept in a wide range of populations, including ones not previously tested? We cannot help but answer this question every time we apply PRO measures in research or practice; it is essential that we put the question to formal testing rather than assume we have the answer.

Acknowledgments This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant (1U01-AR052177). This manuscript was reviewed by PROMIS reviewers, Paul Pilkonis, Arthur Stone, and Paul Crane, before submission for external peer review. Laurie Burke also provided thoughtful comments on an early draft.

Appendix 1: Conference participants

Nancy Amicangelo (patient representative); Dagmar Amtmann, PhD, University of Washington; Meryl Brod, PhD, The Brod Group; Maarten Boers, MD, PhD, Vrije Universiteit; Laurie B. Burke, RPh, MPH OND/CDER/FDA; David Cella, PhD, Northwestern University; Jill Cyranowski, PhD, University of Pittsburgh Medical Center; Susan Czajkowski, PhD, National Institutes of Health; Darren DeWalt, MD, MPH. University of North Carolina; Jacqueline Dunbar-Jacob, PhD, RN, FAAN, University of Pittsburg; Sofia Garcia, PhD, Northwestern University; Richard Gershon, PhD, Northwestern University; Ari Gnanasakthy, PhD, Novartis Pharmaceuticals Corporation; Ron D. Hays, PhD, University of California at Los Angeles; Thomas Hilton, PhD, National Institute on Drug Abuse; Laura Lee Johnson, PhD, National Center for Complementary and Alternative Medicine; Nancy Kline Leidy, PhD, United BioSource Corporation; Eswat Krishnan, MD, Stanford University; William Lenderking, PhD, United BioSource Corporation (ISOQOL Representative); Amye Leong, MBA (Patient Representative); Mary Lynn, PhD, University of North Carolina at Chapel Hill; Susan Magasi, PhD, Northwestern University; Richard Moxley III, MD, University of Rochester Medical Center; Bhash Parasuramen, PhD, AstraZeneca Pharmaceuticals; Donald Patrick, PhD, University of Washington; Charles D. Petrie, PhD, Pfizer Global Development Headquarters (ISOPOR representative); Theodore Pincus, MD, NYU-Hospital for Joint Disease; Louis Quatrano, PhD, National Institute of Child Health and Human Development; Kenneth Rasinski, PhD, University of Chicago; Bryce Reeve, PhD, National Cancer Institute; William Riley, PhD, National Institute of Mental Health; Dennis Revicki, PhD, United BioSource Corporation; Margaret Rothman, PhD, Johnson & Johnson Pharmaceutical Services, LLC; Nan Rothrock, PhD, Northwestern University; Gery Ryan, PhD, Rand Corporation; Jane Scott, PhD, Mapi Values (ISOQOL Representative); Claire Snyder, PhD, Johns Hopkins School of Medicine; Ruth E.K. Stein, MD, AECOM/CHAM; Arthur Stone, PhD, Stony Brook University; Philip Tonkins, MS, DrPH; Peter Tugwell, MD; University of Ottawa; John Ware, QualityMetric Incorporated; Kevin Weinfurt, PhD; Duke University; Gordon Willis, PhD, National Cancer Institute; James P. Witter, MD, PhD NIAMS; Albert Wu, MD, MPH; Susan Yount, PhD, Northwestern University.

References

1. AERA, APA, NCME. (1999). *American Psychological Association*. Washington, DC: Standards for educational and psychological testing.
2. U.S. Department of Health and Human Services. (2006). Food and Drug Administration draft guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register.
3. U.S. Department of Health and Human Services. (2009). Food and Drug Administration guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register.
4. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during the first two years. *Medical Care, 45*(Suppl 1), S3–S11.
5. Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Quality of Life Research, 18*, 1263–1278.
6. Lasch, K., Marquis, P., Vigneux, M., et al. (2010). PRO development: rigorous qualitative research as the crucial foundation. *Quality of Life Research, 19*, 1087–1096.
7. Charmaz, K. (2006). *Constructing grounded theory: A practical guide though qualitative analysis*. Washington, DC: Sage.
8. Charmaz, K. (2003). Grounded theory: Objectivist and constructivist methods. In: G. Lincoln & M. Day (Eds.), *Strategies for qualitative inquiry* (2nd ed.). Thousand Oaks, CA: Sage Publications
9. Strauss, A., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage Publications.
10. Morse, J., Barnett, N., Mayan, M., Olson, K., Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*, Article 2. URL: http//www,ualberta.ca/~ijqm.
11. Bowen, G. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research, 8*, 137–152.
12. Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? *Field Methods, 18*, 59–82.
13. Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L., & Petrie, C. D. (2009). Use of existing patient-reported

outcome (PRO) instruments and their modification: The ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PROTask force report. *Value in Health, 12*, 1075–1083.

14. Triandis, H. (1994). *Culture and social behavior*. New York: McGraw-Hill, Inc.
15. Hays, R. D., & Fayers, P. (2005). Evaluating multi-item scales. In P. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (2nd ed., pp. 41–53). New York: Oxford University Press.
16. Revicki, D. A., Sorensen, S., & Wu, A. W. (1998). Reliability and validity of physical and mental health summary scores from the medical outcomes study HIV health survey. *Medical Care, 36*, 126–137.
17. Joeskog, K. G. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
18. Hays, R. D., Revicki, D. A., & Coyne, K. S. (2005). Application of structural equation modeling to health outcomes research. *Evaluation and the Health Professions, 28*, 295–309.
19. Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
20. Stull, D. E. (2008). Analyzing growth and change: Latent variable growth curve modeling with an application to clinical trials. *Quality of Life Research, 17*, 47–59.
21. Stull, D., Vernon, M. K., Legg, J. C., Viswanathan, H. N., Fairclough, D., & Revicki, D. A. (2010). Use of linear growth curve models for assessing the effects of darbepoetin alpha on hemoglobin and fatigue. *Contemporary Clinical Trials, 31*, 172–179.
22. Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
23. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care, 45*(Suppl 1), S22–S31.
24. DeWalt, D., Rothrock, N. P., Yount, S. P., & Stone, A. A. P. (2007). on behalf of the PCG. Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(Suppl 1), S12–S21.