Volume 10 • Supplement 2 • 2007
VALUE IN HEALTH

Patient-Reported Outcomes: Instrument Development and Selection Issues

Ralph R. Turner, PhD, MPH,[1] Alexandra L. Quittner, PhD,[2] Bhash M. Parasuraman, PhD,[3] Joel D. Kallich, PhD,[4] Charles S. Cleeland, PhD,[5] the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group

[1]Phase V. Technologies Inc., Wellesley Hills, MA, USA; [2]University of Miami, Miami, FL, USA; [3]Health Economics and Outcomes, Astra Zeneca, Wilmington, DE, USA; [4]Global Health Economics and Outcomes Research, Amgen, Thousand Oaks, CA, USA; [5]MD Anderson Center, Houston, TX, USA

ABSTRACT

At its most elemental, patient-reported outcomes (PRO) assessment involves asking the patients questions and evaluating their answers. Instrument developers need to be clear about what they want to know, from whom they want to know it and why, whether what they learned is credible, and whether they can interpret what they learned in the context of the research objectives. Because credible instrument development is neither inexpensive nor technically trivial, researchers must first determine that no available measure meets their research objectives. We suggest that the tasks of either reviewing current instruments or developing new ones originate from the same basic premise: PRO assessment requires a well-articulated conceptual framework. Once defined in the context of the research objectives, the conceptual framework needs to be adapted to the population of interest. We discuss how qualitative methods enrich the conceptual framework and facilitate the technical measurement tasks of item development, testing, and reduction. We recognize that PRO assessment stands at a technological crossroads with the increasingly frequent application of "modern" psychometric methods and discuss how innovations such as item banks and computer-adaptive testing will influence PRO instrument development. Although items are the essential building blocks for instruments, scales are the primary unit of analysis for PRO assessment, and we discuss methods for scoring and combining them. Finally, PRO assessment is meaningless if the key figure chooses not to cooperate. We consider how respondent burden influences the quality of PRO assessment.

Keywords: assessments, instrument development, patient-reported outcomes.

Introduction

Perspective

Health-care stakeholders need assurance that patient-reported outcomes (PRO) data collected from the instrumentation is credible. "Credibility" should be judged by two general criteria: Do the data stem from an instrument that has a theoretically sound conceptual framework? Do the data meet necessary and sufficient psychometric standards?

Credible instrument development is neither inexpensive nor technically trivial. Researchers providing information from new instruments must make the case that none of the available measures meet the research objectives. Current measures must inadequately address the conceptualization of the research question, or the psychometric evidence must be insufficient, or both. This article addresses the key points that investigators wishing to develop new instruments must cover.

Address correspondence to: Ralph R. Turner, Phase V Technologies Inc., 20 Walnut Street, Ste. 110, Wellesley Hills, MA 02481, USA. E-mail: faze5rt@msn.com

10.1111/j.1524-4733.2007.00271.x

Thoroughly addressing the issue regarding existing measures is the critical first step in instrument development. By considering the issues discussed in this article, researchers may conclude that existing instruments do, in fact, meet their research requirements. Alternatively, they may have better evidence to support their decision to invest in new instrument development. In either case, investigators will be in a better position to make an informed choice about how to proceed.

Importance of a Conceptual Framework

Instrument development proceeds from a conceptual framework. Developers need to describe the theory that influenced choices about the constructs measured. Despite a large and growing body of literature [1,2] and a proliferation of PRO instruments (we found more than 2000 PubMed citations for PRO instrument development articles since 1995), no single set of necessary and sufficient PRO concepts is universally accepted. Rothman et al. [3] in this supplement address the issues involved in establishing conceptual foundations. Our focus is on the critical role that a conceptual framework plays in directing PRO instrument development.

Instrument developers must articulate how a particular conceptual framework guided their construct selection, item development, and psychometric testing. Several examples of well-developed theoretical models are available, ranging from the community-centered model that underlies the Medical Outcomes Study Short Form (SF-36) [4] to utility models that specify trade-offs between quantity and quality of life [5–7]. The community-centered theory illustrates how the framework forms the basis for, and predicts the associations among, the relevant constructs. The theory postulates that illness affects a hierarchy of functional domains, with physical functioning at the base. Although the physical effects of disease are central (e.g., affecting mobility, energy, or other domains), effects are also perceived in more distal domains, such as psychological and emotional functioning, and ultimately in the performance of social roles [8].

This theoretical framework leads to a multidimensional measurement model that assesses functioning across a hierarchy of four principal domains: physical, psychological, social, and symptomatic [9]. In this model, physical functioning and symptoms are typically weighted more heavily, and they are expected to be correlated with other domains, such as psychological and role functioning. Furthermore, scores on this measure should be systematically related to indices of morbidity, external measures of psychological functioning, and "real time" measures of daily functioning [10].

Developers need to specify the intended purpose because the characteristics for PRO instruments intended for screening or case-load description are different from those developed to evaluate new medical or behavioral interventions.

Specifying the Target Population

Because PRO measures are designed to reflect the subjective perceptions of the patient, specifying respondent characteristics is a critical requirement for new instrument development. Research indicating differing health-related quality of life (HRQOL) perceptions—between patients, nurses, physicians and spouses, between children and their parents, and between children and health-care professionals [11–13]—underscore the importance of taking respondent characteristics into account in instrument development.

Although patients are the preferred respondents, for some populations a proxy respondent is necessary; some patients are too young or too old to complete the measure, some are too ill, and still others have significant developmental delay or cognitive impairment that precludes them from responding. The conceptual model may call for the perspective of both the patient and proxy respondent; for example, when assessing treatment burden on caregivers [14–17] or spouses [18–21] or when determining the parents' perspective of their child's asthma symptoms.

The underlying framework dictates the manner in which investigators can incorporate key demographic variables into the development process. Although age, sex (and perhaps gender), race and ethnicity, literacy level, and developmental ability should all be taken into account when an instrument is being developed, the approach will differ if the instrument is to be generalizable or focused. Generalizing to well-explicated target populations should be distinguished from generalizing across populations [22]. Both qualitative research and subsequent psychometric testing need to incorporate study designs that allow the developer to detect possible threats to external validity.

Methods for Specifying Domain Content

The theoretical framework defines the context for generating the initial item pool. Developers need to specify the techniques they employed to produce the item pool. Appropriate methods include literature and web-based searches, in-depth interviews and focus groups with patients (and proxies), and interviews with experts in the field.

Literature Review

Developers need to describe the search strategies they selected to review journals and bibliographic databases. Medical databases (e.g., MEDLINE) are obvious first steps, but often PRO instrument developers overlook other disciplines. The American Psychological Association's PsycINFO database [23] covers the literature in psychology psychiatry, nursing, sociology, education, pharmacology, physiology, linguistics, and other areas. The database encompasses references and abstracts to more than 1300 journals and dissertations in more than 30 languages and to books in the English language.

The International Bibliography of Social Sciences [24] indexes the information contained in more than 2600 social sciences journals and 6000 books each year. Coverage includes both core and specialized material from more than 100 countries in more than 90 languages. Approximately 70% of the records are in English, and articles in other languages are displayed with both the original language title as well as an English translation. EMBASE: Rehabilitation and Physical Medicine is an international biomedical and pharmacological database containing citations and abstracts to journal literature on rehabilitation of both physical and mental disorders using physiotherapy and other therapies [25]. EMBASE contains more than 10 million records from 1974 to present, with 500,000 citations and abstracts added annually. EMBASE features comprehensive coverage of drug research, pharmacology, pharmacy, pharmacoeco-

nomics, pharmaceutics and toxicology, human medicine (clinical and experimental), basic biological research, health policy and management, public, occupational and environmental health, substance dependence and abuse, psychiatry, forensic science, and biomedical engineering and instrumentation.

Special consideration needs to be given to the potential sources of bias that might emerge from the literature. For instance, an instrument that is designed to measure symptoms in chronic obstructive pulmonary disease (COPD) [26] may not be the right instrument if the researcher wants to measure loss of functionality because of COPD symptoms. Additional biases may be introduced with the particular choice of Medical Subject Heading terms used while doing this literature search. Information gained during this step should be used to aid in the development of a qualitative discussion guide for preliminary work in the construction of a new measure.

Other Instrument Review

Existing instruments that address the same or related areas of PRO assessment should be identified and reviewed. Existing instruments may not be measuring specific domains of interest (e.g., rheumatoid arthritis-related fatigue). The degree to which some well-established instruments are accepted for certain applications may change. For example, the symptom-specific Baseline and Transition Dyspnea Indices have an extensive publication history but faced some criticism during the approval of tiotropium [27].

Several comprehensive instrument review resources are available to developers. The Mental Measurements Yearbook Test Reviews On-line [28] covers more than 2000 commercially available educational, personality, aptitude, neuropsychological, achievement, and intelligence tests. The Behavioral Tests and Measures in the Health Sciences Resource Guide [29] describes major resources used to locate materials on psychological or behavioral questionnaires, surveys, measures, tools, scales, and instruments as related to the health sciences. Health and Psychosocial Instruments (HaPI) [30] provides ready access to information on measurement instruments, including questionnaires, interview schedules, checklists, index measures, coding schemes/manuals, rating scales, projective techniques, vignettes/scenarios, and tests in the health fields, psychosocial sciences, organizational behavior, and library and information science.

Qualitative Assessment in Instrument Development and Selection

Qualitative research is an important component in PRO instrument development, and may also be a critical component of selection among existing instruments. Qualitative research needs to be approached with the same sense of rigor that accompanies quantitative assessment. The use of a semistructured discussion guide, developed from the conceptual framework and literature review, is essential. The focus group should be large enough to generate diverse viewpoints, but small enough to be manageable; Krueger [31] recommends seven to 10 people per focus group. To be effective, a trained and experienced professional should moderate focus groups, and developers should convene at least three separate groups. PRO researchers need to be mindful of several issues that can confound the content development during focus groups. These include: 1) bias inherent in the development of discussion guide; 2) bias introduced by the discussion leader; 3) dominance effects; and 4) lack of experience of group to the concept under investigation.

In-depth Interviews

We endorse in-depth interviews and cognitive debriefing as critical methods in PRO instrument selection and development. Such focused, one-on-one conversations with relevant individuals (patients from the assessment target populations) conducted by trained staff provide valuable insight into the respondent's perception, feelings, and perspectives of the disease condition. Key to success is anticipating and organizing the issues that are to be explored. These interviews provide an opportunity to follow up on questions and probe for deeper meaning and understanding of the responses. Instrument developers need to be aware of, and address, issues that may affect the interview content. The experience of the first few subjects could bias the line of questioning that the interviewer takes with subsequent subjects. In-depth interviewing is time-consuming and resources need to be planned accordingly. Small samples are not likely to be fully representative of the target population. Patients should be considered the primary source of information, especially when the PRO relates to concepts such as patient satisfaction and psychosocial function. In-depth interviews are critical to both item generation and scale construction.

Cognitive debriefing, a less time-intensive step, consists of a structured interview that asks patients how well they understand items of new or existing scales, how comfortable they are with answering the items, and how well the items reflect their concerns with their disease or treatment. Cognitive debriefing may be an iterative process in scale development. Cognitive debriefing should also be helpful in evaluating a choice among existing scales where the original scale development did not include substantial patient input.

Target Population Surveys

As an effective tool for obtaining patient-level feedback of specific content, target population surveys present a highly structured series of written questions that are administered to a large number of persons

with a specific disease or condition. They are useful to describe populations, to assess the prevalence of behavior and knowledge of populations, and to get a preliminary quantitative assessment of the concepts under study. Survey design is important to maximize the resources invested. Sample sizes depend on the objectives, and investigators can estimate them from either preliminary data or the literature. Using dichotomous data (or creating them from polytomous data), sample size requirements for the desired level of precision can be estimated from the proportional response distribution and the standard error: $N = P(1)P(0)/\text{Std Error}^2$.

Content Analysis of Available Data Sources

Online "blogs" or chat rooms, and other data sources provide a readily available and inexpensive source for content analysis. For example, the site Data on the Net (University of California, San Diego) links to 156 different sites that provide social science data [32]. Most diseases and medical conditions have user groups, allowing for sampling content from target populations. Content can be downloaded and analyzed with qualitative analysis software (some 20 different programs are available, including freeware from the Centers for Disease Control and Prevention, http://www.cdc.gov/HIV/software/ez-text.htm) to provide support for domain specification. Although information gained can be used in the development of PRO measures, further work is needed to determine what bias such self-selected samples are likely to contain.

Expert Panels

Both informal and formal consensus methods ought to be utilized during the PRO instrument development process. Informal methods—"Three Smart People in a Room"—often yield valuable guidance from a small group of content experts. This information can be used in more formal consensus methods; for example, Delphi or Nominal Group techniques, where formal inquiry is interspersed with controlled feedback [33–36]. In either case, the list of items generated during the qualitative interview process should be reviewed by clinicians experienced in treating or managing patients from the disease area in question to ensure: 1) coverage of signs, symptoms, and other issues most often reported by patients; and 2) use of terminology used by patients in the clinic setting. Adding this step to the development of PRO items addresses the issue of content validity.

Item Selection: The Construct–Respondent Interface

Inexperienced instrument developers are tempted to begin with the items. Previous sections in this article, as well as other articles in this supplement (e.g., Rothman et al. [3]), counsel against this starting point; instead, we argue for a well-founded theoretical basis and thorough familiarity with the relevant work in the area. Only then should an item-pool be developed and tested.

The conceptual framework determines the item content; item formatting decisions depend on the intended uses for the instrument. Kirshner and Guyatt's taxonomy [37] of descriptive, evaluative, and predictive measures is a useful framework. The technical issues involved in designing and testing items are covered thoroughly in several psychometric texts (cf. [38–40]).

Item Response Theory Assessment

Patient-reported outcomes assessment is at a technological crossroads. As the patient's role in health-care decisions becomes more prominent, the technology available to convey his or her state of health and well-being is reaching new levels of sophistication. For a decade or so, social science and educational researchers have applied the "modern" psychometric techniques of item banking, item response theory (IRT), and computer-adaptive testing (CAT) to assessment development [41], but only relatively recently and with an accelerating pace have these methods made their way onto the PRO assessment stage. In June 2004, the National Cancer Institute and the Drug Information Association sponsored a conference entitled "Advances in Health Outcomes Measurement: Exploring the Current State and Future Applications of Item Response Theory, Item Banks, and Computerized-Adaptive Testing," which represented an important opportunity for PRO researchers to discuss this advancement in technology [42].

These developments have important implications for PRO instrument design. Most applications continue to rely on static assessment. Researchers have to balance coverage, precision, specificity, and respondent burden. IRT methods have the potential of reducing the need for additional static instruments to plug gaps in the PRO assessment spectrum. Several initiatives, notably the Patient-Reported Outcomes Measurement Information System project, are addressing the issue of developing calibrated item banks (http://www.nihpromis.org). By knowing the precise location of the item "difficulty" (or severity) with respect to the underlying latent PRO domain, researchers are able to tailor instruments that optimize the coverage and precision with the fewest number of items, thereby addressing the respondent burden issue as well. Although the goal of a nationally normed calibrated item bank is still over the horizon, "local" item banks calibrated to specific diseases provide an important interim step in advancing PRO instrument development and application [43].

Computer-adaptive testing has tremendous potential for yielding precise PRO assessment quickly and with significantly reduced respondent burden. Although CAT provides technical solutions to many PRO assessment problems, it is not yet suited to all applications [44].

Scale Development and Scoring

The item is the unit of analysis in psychometrics, the scale is most often the unit of analysis in PRO assessment. Although the assumptions underlying the properties of a scale differ between classical and modern measurement theory, the objective is the same: the scale is the building block for the conceptual framework.

Scales provide the basis for generalizing the results beyond the specific content of an item, from "I can climb one flight of stairs" to mobility, or "I feel downhearted and blue" to depression. These conceptual building blocks require evidence that they are the optimal combination of items that conveys the conceptual meaning, including internal consistency, convergent and divergent validity, differential item functioning, and the like. Once the scale has been established, the individual items are often ignored and the analysis and interpretation center around anxiety, pain, sexual functioning, and the like.

Scale scoring lies at the heart of the interpretation debate. The most frequently asked question of PRO assessors is "What do these scores mean?" There are two facets of equal importance to this question: frame of reference and clinical meaning.

The first important facet involves a clear frame of reference. PRO assessment does not (yet) have a standard reference. One advantage of the SF-36 is the existence of national norms. Although many applications do not involve direct comparisons with national norms (examining treatment differences between two agents, for example), consumers of PRO information often seek to place the results in some larger frame of reference.

The increasingly common practice of converting all scale scores to a 0–100 scale is an attempt to address this issue. One solution, widely used in educational research and being employed by some PRO researchers, is to convert the scales to T-scores, centering the distribution at a common mean (e.g., 50) and establishing a common standard deviation (e.g., 10). Additional items of varying "difficulty" or severity can be included in the scale without affecting the scaling.

The second facet involves "clinical significance" or "minimally important [clinical] differences." This issue represents a considerable barrier to widespread acceptance of PRO information in health-care research. The issue has been given careful thought, notably the "Symposium on the Clinical Significance of Quality-of-Life Measures in Cancer Patients" [45]. Instrument developers need to include information on minimally important difference research in the PRO instrument dossier, including the selection of the approach (anchor- or distribution-based), as well as the study design and results.

Summary Scale Development and Scoring

Summary scales, also known as summated rating scales or Likert summated scales, are a set of questions (items), all of which are considered to be of approximately equal value and to which subjects respond with degrees of agreement or disagreement (intensity). The score on the questions are either summed or averaged to yield an individual score per subject [38,46,47]. This idea has been advanced further to create an even higher-level summary that is a summated scale of two or more summary scales [48]. Investigators should develop and use summary scales when they need to limit the number of outcomes being analyzed (i.e., the multiple comparison issue), and when the effect being investigated is considered to be extremely general in nature and expected to affect several different domains of quality of life.

The primary goals of creating summary scales are twofold: first, to place an individual somewhere on a continuum of the construct that is being measured (e.g., "physical functioning"); and second, to reduce the number of statistical comparisons required to analyze the construct without loss of information [47,48]. Scales often comprise many items to increase the reliability, precision, and validity of the measurement of the construct. Often, factor analysis or principal component analysis are used to evaluate if several items or scales are measuring the same underlying construct and could be grouped together. As Ware has pointed out, psychometrically based summary measures or aggregated health measures often lead to the same conclusion [48].

Summated scales belong to the classical psychometric model of parallel test items. In this approach, developers make three main assumptions: 1) each item is a measure of the same underlying construct; 2) the items have similar statistical properties; and 3) the items can be easily combined. Adaptive testing may allow researchers to collect the necessary detail for a subpopulation of interest, while collecting only key items that are necessary for general population norms are collected from all subjects. This new approach achieves the twin goals of minimizing respondent burden and allowing for complete collection of the attributes that characterize the subpopulation's health status.

Necessary and Sufficient Evidence for a "Summary Scale" in Static Testing

The analytic evidence that investigators typically consider for construction of a summary scale entails either

factor analysis or principal component analysis. Items should have the same possible range of score values; otherwise, variance because of response sets can complicate the interpretation of these analyses. The pattern of correlations observed between items should support the conclusion that each item is approximately an equally good indicator of the same underlying construct. In other words, the summary scale should account for a substantial amount of the variance without single items not in the summary scale contributing to the variance explained. Confirmation of the variance-explained results in different patient populations provides further evidence supporting the validity of the construction of the summary scale. Thus, additional factor analysis or principal component analysis should be done in other surveys and should yield the same amount of explained variance by the summary scales.

Nevertheless, summary scales constructed of "symptoms" should be approached with extreme caution. By their very nature, different symptoms have differential intensities of effect on patients and their quality of life. Without weighting the symptoms by patients' ratings of importance, one can commit a grievous error of minimizing a symptom that has an overwhelming impact on a patient's quality of life.

Interpretability of Scores

Investigators often find it difficult to determine whether the simplicity and ease of a summary measure in statistical analysis obscures important change in a particular domain or area of health-related quality of life. This problem of interpretation becomes more complex if population issues are at stake. A case in point may be whether the elderly differ dramatically from the general population on a particular health domain or item but not on a summary scale of general health.

Consideration of Consumers of Scores: Patients, Providers, Caregivers, and Regulators

Summary scales are a psychometrically sound approach that provides consumers with an easy method of scoring and a readily interpretable score that has sensitivity and specificity for a particular domain. Utility scores or indices, by contrast, are measures that are aggregated without consideration of the underlying domains. They have been shown to have floor and ceiling effects and low correlations with other general quality-of-life measures.

Administrative Burden

Investigators need to take account of two main determinants of respondent burden. The first is length of assessment. A second determinant is ease of responding and patient comprehension of the assessment task, especially in the context of the trial. Trialists need to consider whether the assessment "makes sense" to the patients in terms of what question the trial is trying to answer.

The most obvious reason to minimize respondent burden is for patient comfort and convenience. From the trial perspective, reducing respondent burden is a major factor in reducing missing data. Respondent burden is situational. Patients who have long waiting periods during their clinic visit may be more willing to spend longer periods completing an assessment. Doing assessments electronically (computer or interactive voice response systems) may allow patients to complete longer assessments at home, without the pressures that accompany clinic visits (multiply scheduled consultations, concern about transport to and from the clinic, concerns about childcare or returning to a job). Finally, some patients are just too ill to complete more than a very short assessment.

Reduction of respondent burden should be a major consideration in trial design. Assessors need to consider which instruments can answer the trial questions most parsimoniously. If more than one instrument needs to be used, then the strategy extends to consideration of the minimal set of instruments in the battery. If sufficient data exist, a psychometric approach (using variants of factor analysis) can be used to assure that only the most trial-relevant questions or measures are used. Finally, not all assessment instruments may need to be given at each assessment point. For example, in a symptom reduction trial, simple ratings of symptom severity are often used at more frequent points than measures of function, global quality of life, or satisfaction with treatment.

Respondent burden associated with an assessment battery is multiplied by the number of assessment points required by the study design. The more assessment points that are required (longitudinal assessment), the more compact the assessment needs to be. Often, two important aspects of the trial may help solve these design questions: 1) how quickly is the agent of interest expected to have its first clinically meaningful effects; and 2) when is its maximum benefit expected to be achieved (latency of effect). Longitudinal assessment is required to answer these questions. The time period(s) in question may be hours, days, or weeks, or in some cases (relief of traumatic or cancer pain crisis), even minutes. The severity of the condition and its implications for the patient drives the need for rapidly acting interventions. Severe depression is an example. Another example is the treatment of bone pain in advanced cancer, which can be addressed by several different treatment approaches (radiation, surgery, bisphophonates); thus, the latency of effect onset (days versus weeks) becomes critically important in deciding which treatment to use for patients with limited lifespan.

Longitudinal assessment is also required to determine the duration of effects. Many agents, especially agents directed at symptom reduction, work "for a while." Their beneficial effects may taper off for several reasons (development of drug resistance, increasing tolerance to the agent, etc.). Patients may also recalibrate the trade-off between the positive benefits of the agent and its negative side effects or the inconvenience of taking the agent. Only a subset of agents that are effective will be so over time and will continue to be used by the target population. To be able to describe the duration of effects, depending on the agent to be evaluated, will require evaluation over a period of weeks or months. Here is where missing data become an appreciable problem, mandating a very brief assessment and substantial planning on how this assessment is to be collected. Often, research nurses or data managers collect such data by telephone, especially when patients/clinical condition no longer requires routine clinic visits. Electronically assisted assessments (computer or interactive voice response systems) may help in the collection of these long-term data.

Conclusions

Patient-reported outcomes assessment builds on a long and empirically based approach to measurement. It is not "new," nor "soft," nor "subjective" in the negative sense with which that term is used to describe the data. Rather, effective PRO assessment stems from a conceptual framework that explains and predicts patient behavior. Instrument development is neither trivial nor inexpensive. Done correctly, it involves theoretical explication, qualitative exploration, quantitative confirmation, and psychometric support. Instrument developers need to be clear about what they want to know, from whom they want to know it and why, if what they learned is credible, and whether they can interpret what they learned in the context of the research objectives. High-quality PRO assessment data will take on increasing importance as they become more central in the health-care delivery arena.

Source of financial support: Funding for the meeting was provided by the Mayo Foundation in the form of unrestricted educational grants; North Central Cancer Treatment Group (NCCTG) (CA25224-27) and Mayo Comprehensive Cancer Center grants (CA15083-32).

References

1 Fayers P, Machin D. Quality of Life: Assessment, Analysis, and Interpretation. Chichester: John Wiley & Sons, 2000.
2 Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ (Clinical Research ed.) 2002;324:1417.
3 Rothman ML, Beltran P, Cappelleri JC, et al. Patient-reported outcomes: conceptual issues. Value Health 2007;10(Suppl. 2):S66–75.
4 Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36™). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
5 Kaplan RM. Health outcome models for policy analysis. Health Psychol 1989;8:723–35.
6 Torrance G. Toward a utility theory foundation for health status index models. Health Serv Res 1976;11:349–69.
7 Torrance G. Measurement of health state utilities for economic appraisal. J Health Econ 1986;5:1–30.
8 Alonso J, Ferrer M, Gandek B, et al. Health-related quality of life associated with chronic conditions in eight countries: results from the International Quality of Life Assessment (IQOLA) Project. Qual Life Res 2004;13:283–98.
9 Diamond R, Becker M. The Wisconsin Quality of Life Index: a multidimensional model for measuring quality of life. J Clin Psych 1999;60(Suppl. 3):S29–31.
10 Modi A, Quittner A. Validation of a disease-specific measure of health-related quality of life for children with cystic fibrosis. J Ped Psychol 2003;28:535–46.
11 Chang P, Yeh C. Agreement between child self-report and parent proxy-report to evaluate quality of life in children with cancer. Psycho-Oncology 2005;14:125–34.
12 Hickey A, Barker M, Mcgee H, O'Boyle C. Measuring health-related quality of life in older patient populations. Pharmacoecon 2005;23:971–93.
13 Jokovic A, Locker D, Guyatt G. How well do parents know their children? Implications for proxy reporting of child health-related quality of life. Qual Life Res 2004;13:1297–307.
14 Canam C, Acorn S. Quality of life for family caregivers of people with chronic health problems. Rehabil Nurs 1999;24:192–6.
15 Le T, Leis A, Pahwa P, et al. Quality-of-life issues in patients with ovarian cancer and their caregivers: a review. Obstet Gynecol Surv 2003;58:749–58.
16 White CL, Lauzon S, Yaffe MJ, Wood-Dauphinee S. Toward a model of quality of life for family caregivers of stroke survivors. Qual Life Res 2004;13:625–38.
17 Visser-Meily JM, Post MW, Riphagen II, Lindeman E. Measures used to assess burden among caregivers of stroke patients: a review. Clin Rehabil 2004;18:601–23.
18 Rees J, O'Boyle C, Macdonagh R. Quality of life: impact of chronic illness on the partner. J R Soc Med 2001;94:563–6.
19 Sneeuw K, Sprangers M, Aaronson N. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease. J Clin Epidemiol 2002;55:1130–43.
20 Testa M, Hollenberg N, Anderson R, William G. Assessment of quality of life by patient and spouse during antihypertensive therapy with atenolol and

nifedipine gastrointestinal therapeutic system. Am J Hypertens 1991;4:363–73.
21 Testa M. Parallel perspectives on quality of life during antihypertensive therapy: impact of responder, survey environment, and questionnaire structure. J Cardiovasc Pharm 1993;21(Suppl. 2):S18–25.
22 Cook T, Campbell D. Quasi-experimentation: design and analysis issues for field settings. Boston, MA: Houghton Mifflin Co., 1979.
23 Psycinfo. American Psychological Association, Washington, DC. 2006. Available from: http://www.apa.org/psycinfo/ [Accessed August 27, 2007].
24 International Bibliography of Social Sciences, Economic and Social Research Council, and the London School of Economics. British Library of Political and Economic Science of London School of Economics and Political Science, London, England. 2005. Available from: http://www.lse.ac.uk/collections/IBSS/ [Accessed March 14, 2007].
25 EMBASE. Elsevier B.V., New York, NY. 2006. Available from: http://www.embase.com [Accessed March 14, 2007].
26 Witek TJ Jr. Validation of the BDI/TDI clinical thresholds using clinical trials data. Eur Res Rev 2002; 12:43–55.
27 Casaburi R, Mahler D, Jones P, et al. A long-term evaluation of once-daily inhaled tiotropium in chronic obstructive pulmonary disease. Eur Res J 2002;19: 217–24.
28 Mental Measurements Yearbook Test Reviews On-Line. Buros Institute of Mental Measurements, University of Nebraska-Lincoln, Lincoln, NE. 2004. Available from: http://www.unl.edu/buros/bimm/index.html [Accessed March 14, 2007].
29 Behavioral Tests & Measures in the Health Sciences Resource Guide. Harvey Cushing/John Hay Whitney Medical Library, Yale University School of Medicine. Available from: http://info.med.yale.edu/library/reference/publications/tests.html #Databases [Accessed March 14, 2007].
30 Health and Psychosocial Instruments (HaPI). Ovid Technologies Inc., New York, NY. 2002. Available from: http://www.ovid.com/site/products/ovidguide/hapidb.htm#geninfo [Accessed March 14, 2007].
31 Krueger RA. Focus Groups: A Practical Guide for Applied Research. Thousand Oaks, CA: Sage, 1994.
32 Data on the Net. University of California, San Diego. Available from: http://3stages.org/idata/ [Accessed March 14, 2007].
33 Alder M, Ziglio E, eds. Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health. New York: Jessica Kingsley, 1996.
34 Delbecq AL, Vandeven AH. A group process model for problem identification and program planning. J Appl Beh Sci 1971;7:466–92.
35 Linstone HA, Turoff M, eds. The Delphi method: techniques and applications. Available from: http://www.is.njit.edu/pubs/delphibook/ [Accessed July 18, 2006].
36 Delbecq AL, VandeVen AH, Gustafson DH. Group Techniques for Program Planners. Glenview, IL: Scott Foresman, 1975.
37 Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985; 38:27–36.
38 Nunnally J, Bernstein I. Psychometric Theory (3rd ed.). New York: McGraw-Hill Series in Psychology, 1994.
39 Kline P. A Psychometric Primer. London: Free Association Books, 2000.
40 Streiner D, Norman G. Health Measurement Scales: A Practical Guide to Their Development and Use (3rd ed.). Oxford: Oxford University Press, 2003.
41 Bjorner JB, Ware JE. Computer software resources for Rasch or IRT models. Medical Outcomes Trust Bulletin, April 1998.
42 Reeve BB. Special issues for building computerized-adaptive tests for measuring patient-reported outcomes: the National Institute of Health's investment in new technology. Med Care 2006;44(11 Suppl. 3):S1987–204.
43 Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. Qual Life Res 2007; 16:109–19.
44 Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005;28:254–82.
45 Sloan J, Cella D, Frost M, et al. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. Mayo Clin Proc 2002;77: 367–70.
46 Kerlinger FN, Lee HB. Foundations of Behavioral Research (4th ed.). Fort Worth, TX: Harcourt, 2000.
47 Fayers P, Curran D, Machin D. Incomplete quality of life data in randomized: missing items. Stat Med 1998;17:679–96.
48 Mchorney C, Ware J, Raczek A. The MOS, 36-item short-form health survey (SF-36™): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31: 247–63.