Methodological Issues In The Selection, Administration And Use Of Patient-Reported Outcomes In Performance Measurement In Health Care Settings

Final manuscript, Sept.28, 2012

David Cella, Ph.D., Elizabeth A. Hahn, M.A., Sally E. Jensen, Ph.D., Zeeshan Butt, Ph.D., Cindy J. Nowinski, M.D., Ph.D., Nan Rothrock, Ph.D.

Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University
710 N. Lake Shore Drive
Abbott Hall, Suite 729
Chicago, IL 60611

Phone: 312-503-1725

Email: d-cella@northwestern.edu

The authors thank Kathleen Swantek, MLIS, for assistance with reference management, and we thank the following for their helpful comments on the content of this paper: Karen Adams, Ph.D., MT; Ethan Basch, MD, MSc; Victor Chang, MD; Stephan Fihn, MD, MPH; Floyd Jackson Fowler, Ph.D.; Lewis Kazis, Sc.D.; Kathleen Lohr, Ph.D.; Jennifer Moore; Eugene Nelson, DSc, MPH; Kenneth Ottenbacher, Ph.D., OTR; Karen Pace, Ph.D., MSN; and Mary Tinetti, MD.

I. Introduction

The increasing integration of health care delivery systems provides an opportunity to manage the entire patient-focused episode of care[1] and to assess the impact of care on patient outcomes, including patient-reported outcomes. The National Quality Forum (NQF) commissioned this paper as part of an initiative to find patient-reported outcomes that might be considered along with the tools by which data on them are collected as a new dimension in the types of performance measures NQF endorses. This paper reviews issues to consider when evaluating patient-reported outcomes (PROs)[2] as candidate performance measures in health care settings. Consistent with the NQF perspective, we distinguish patient reported outcome measures (PROs or PROMs), from patient reported outcome – performance measures (PRO-PMs), with the latter being that which NQF endorses. This paper addresses the PROs that are likely to be used to inform PRO-PMs, understanding that there is a companion paper that addresses validity concerns related to PRO-PMs.

PROs are defined here as "any report of the status of a patient's health condition, health behavior, or experience with health care that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" (See Table 1). In other words, PRO tools measure what patients are able to do and how they feel by direct, unfiltered inquiry. A large literature supports the use of PROs; it provides cogent evidence suggesting that clinical providers are limited in accurately estimating outcomes for patients.[3-7] PRO tools enable clinicians, patients and families, and others to assess patient–reported health status domains (e.g., health status; physical, mental, and social functioning; health behavior; experience with health care). A wide variety of patient-level instruments to measure PROs have been used for clinical research purposes and to guide clinical care; many have been evaluated and catalogued in the work conducted by the NIH Patient-Reported Outcomes Measurement Information System (PROMIS®; www.nihpromis.org) cooperative group. The PROMIS® system itself has not yet been used for assessing performance; however, components of it have been used in the past. Two major challenges to using PROs for purposes of accountability and performance improvement must be addressed. First, they have not been widely adopted in clinical use; thus, they are unfamiliar to many health care professionals, payers, and others in health care systems. Second, little is known about the best set of responsive questions to aggregate for the purpose of measuring *performance* of the health care entity.

Many in the health sector are showing increasing interest in moving toward use of PROs for these clinical, quality improvement, and accountability applications. Foundational work still needs to address methodological and data challenges. Efforts are currently underway (mid-2012) to develop and test mechanisms for collecting patient-reported data, so this is an opportune time to consider methodological issues in some depth. These issues include collection of PRO data in the clinical environment and aggregation of data to assess organization- and clinician- level performance.

The purpose of this white paper is to address the major methodological issues related to the selection, administration and use of PROs for individual patients in clinical practice settings. This information will inform the selection of PROs as candidate measures for use in performance assessment and related applications. This paper also identifies best practices in identifying and using PROs in performance measures. A separate white paper will outline the path to developing reliable and valid performance measures eligible for NQF endorsement that various provider groups, regulatory agencies, payers and insurers, and others can use for accountability and quality improvement activities (i.e., PRO-PMs).

Table 1. Definitions and key concepts that are central to the purpose of this paper

Patient: The term used when a person is receiving health care services, or when using long-term health care support services.
Patient-reported outcome measure (PRO or PROM): Any report of the status of a patient's health condition, health behavior, or experience with health care that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. Ideally this is measured using a standardized tool.
Performance measure: Numeric quantification of health care quality for a designated accountable health care entity, such as hospital, health plan, nursing home, clinician, etc.
PRO-based performance measure (PRO-PM): A performance measure that is based on patient-reported outcome data aggregated for an accountable health care entity (e.g., percentage of patients in an accountable care organization with an improved depression score as measured by a standardized tool).
e-health[8,9]**:** Health-related Internet applications that deliver a range of content, connectivity and clinical care. Examples include: online formularies, prescription refills, test results, physician-patient communication.
Patient-Centered Outcomes Research (PCOR): Integration of patient perspectives and experiences with clinical and biological data collected from the patient to evaluate the safety and efficacy of an intervention (www.pcori.org).
Reliability: The extent to which a scale or measure yields reproducible and consistent results.[10]
Validity: The extent to which an instrument measures what it is intended to measure and that it can be useful for its intended purpose.[10]

II. Types of Patient-Reported Outcomes

PROs can be used to assess a wide variety of health-relevant concepts, including health-related quality of life, functional status, symptoms and symptom burden, health behaviors, and the patient's health care experience. These concepts are neither mutually exclusive nor exhaustive. Table 2 summarizes the main characteristics of these types of PROs.

Table 2. Main Characteristics of Patient-Reported Outcomes

PRO Category	Main Characteristics	Strengths	Limitations
Health-Related Quality of Life	• Multi-dimensional • Can be generic or condition-specific	• Global summary of well-being	• HRQL may not be considered a sufficiently specific construct
Functional Status	• Address ability to perform specific activities	• Can be used in addition to performance-based measures of function	• Self-reported capability and actual performance of activities may vary
Symptoms and Symptom Burden	• Specific to type of symptom of interest • May identify symptoms not otherwise captured by medical work-up	• Best assessed through self-report	• May fail to capture general, global aspects of well-being considered important to patients
Health Behaviors	• Specific to type of behavior • Typically measures frequency of behavior	• Targeted to specific behavior categories	• Validity may be affected by social desirability • May be potential patient discomfort in reporting socially undesirable behaviors
Patient Experience	• Satisfaction with health care delivery, treatment recommendations, and medications (or other therapies) • Actual experiences with health care services	• Essential component of patient-centered care • Valued by patients, families and policymakers	• May be a complex, multidimensional construct • Confidentiality is required to ensure patient comfort in disclosing negative experiences

PRO Category	Main Characteristics	Strengths	Limitations
	• Patient activation	• Related to treatment adherence • Related to health behaviors and health outcomes	• Insufficient evidence that activation enhances health care decision-making

Health-Related Quality of Life

One class of PRO measures health-related quality of life (HRQL). HRQL is a multi-dimensional[11] construct encompassing physical, social, and emotional well-being associated with illness and its treatment.[12] Different types of HRQL measures[13,14] are useful for different purposes.[15] Numerous generic health status measures, such as the Medical Outcomes Study Short Form SF-36 and related measures, and the Sickness Impact Profile are classic examples.[16-19] This type of HRQL PRO is useful in assessing both individuals with and without a health condition. Such data allows researchers, clinicians and others to compare groups with and without a specific condition and to estimate population norms. A health utility or preference measure is also not disease-specific. It provides a score ranging from 0 (death) to 1 (perfect health) that represents the value that a patient places on his or her own health.[20] Experts can use scores from these types of measures to calculate quality-adjusted life years or compare information to population norms.

Many PROs are intended for use in populations with chronic illness.[21-23] Recently, the Patient-Reported Outcome Measurement Information System (PROMIS®) has developed a considerable number of PROs in physical, mental, and social health for adults and pediatric samples with chronic conditions.[24,25] Neuro-QOL is another measurement effort focused on capturing important areas of functioning and well-being in neurologic diseases.[26] Each of these measurement efforts does not reference a specific disease in the items; thus, they permit comparisons across conditions. Other PROs are targeted on a specific disease (e.g., spinal cord injury) or treatment (e.g., chemotherapy).[27,28] Often these instruments are developed so that investigators can demonstrate responsiveness to treatment in a clinical trial rather than cmpare data to population norms or information on other conditions.[29] Condition-specific PROs often provide additional, complementary information about a patient's HRQL. [22,30-32]

Functional Status

Another type of PRO is a functional status measure. Functional status refers to a patient's ability to perform both basic and more advanced (instrumental) activities of daily life.[33] Examples of functional status include physical function, cognitive function and sexual function. As with HRQL instruments, a large number of functional status measure exist, but they vary widely in quality.[34] Some may address a very specific type of function (e.g., Upper Limb Functional Index) or, be developed for use in a specific disease population (e.g., patients with multiple sclerosis); others may be appropriate for use across chronic conditions.[35-41]

Symptoms and Symptom Burden

Symptoms such as fatigue and pain intensity are key domains for PRO measures. Symptoms are typically negative; their presence and intensity are best assessed through patient

report.[42] Scales are used to characterize the severity of the symptoms. The impact of symptoms such as the degree to which pain interferes with usual functioning, is also a common focus of PROs. Symptom burden captures the combination of both symptom severity and impact experienced with a specific disease or treatment.[42] Common symptom and symptom burden measures include the Functional Assessment of Chronic Illness Therapy – Fatigue scale and disease-focused symptom indices such as the recent NCCN symptom indexes for various cancer types or a COPD dyspnea-specific instrument.[43,44] The PROMIS® initiative developed the PROMIS® Pain Interference measure which quantifies the impact of pain on functioning.[45]

Health Behaviors

Yet another category of PROs assesses health behaviors. Although health behaviors may be considered predictors of health outcomes, they are also health outcomes in their own right in the sense that health care interventions can have an impact on them. Information from health behavior PROs serves several important clinical purposes. Health behavior PROs can be used to monitor risk behaviors with potentially deleterious health consequences. This information enables clinicians to identify areas for risk reduction and health promotion interventions among their patients. Health behavior PROs can also be used to assess patients' response to health promotion interventions and to monitor health behaviors over time.

Health risk assessments (HRAs) provide an illustrative example of how health behavior PROs can be incorporated into health promotion and disease prevention programs. Defined by the Centers for Disease Control (CDC) as tools to measure individual health, HRAs may consist of clinical examination or laboratory test results as well as health behavior PROs.[46]. A recent AHRQ report identified three key components in the process of implementing HRAs in health promotion: (1) provision of patient self-reported information to identify risk factors for disease, (2) provision of individualized health-specific feedback to patients based upon the information they reported, and (3) provision of at least one health promotion recommendation or intervention.[47] Although HRAs have been implemented in community settings, universities, and health maintenance organizations, they have been most commonly implemented in workplace settings. [47] An extensive review of HRA program outcomes concluded that, in many cases, the implementation of HRA programs improved health behaviors and intermediate health outcomes (e.g., blood pressure); however, the evidence did not demonstrate whether the implementation of HRAs affected disease incidence or health outcomes over the medium to long term. [47]

As the emphasis on the importance of health behaviors has increased, so has the number of available PROs developed to assess health behaviors across multiple domains. Health behavior PROs may assess general health by measuring risk factors without a focus on a specific disease or behavioral category. The Personal Wellness Profile™[48] and the *Insight*® Health Risk Appraisal Survey[49] are examples of health behavior PROs measuring multiple risk factors that have been certified by the National Committee for Quality Assurance. In addition, several large-scale health behavior assessment systems provide additional context for the use of general health behavior PROs. The Behavioral Risk Factor Surveillance System (BRFSS) was created in 1984 by the CDC as a state-based system utilizing a standardized questionnaire to measure health risk and health promotion behaviors, such as health awareness, tobacco use, consumption of fruits and vegetables, physical activity, seatbelt use, immunization, and alcohol consumption.[50] The National Health and Nutrition Examination Survey (NHANES) constitutes another large-scale implementation of health behavior PROs. NHANEs was established by the CDC in the 1960's and includes health behavior surveys in addition to clinical examinations to assess health status at the population level.[51] The health behavior survey portion of NHANES assesses a range of health risk and health promotion behaviors, including smoking, drug use,

alcohol use, sexual practices, physical activity, dietary intake, and reproductive health practices.[51]

Health behavior PROs can also assess risk factors associated with specific diseases (e.g., smoking) or risk factors associated with specific behavioral categories (e.g., physical activity, seatbelt use, food consumption). Examples include: the health risk survey, an interactive computer-based health risk survey assessing alcohol consumption and smoking;[52] the CAGE-Adapted to Include Drugs (CAGE-AID), a self-reported screening measure of substance use disorder among treatment-seeking adolescents.[53] A subset of health behavior PROs assesses health-promoting behaviors. "Starting the conversation," a brief measure of dietary intake,[54] "Exercise as the fifth vital sign," a brief measure of physical activity,[55] School Health Action, Planning and Evaluation System (SHAPES), a school-based self-report physical activity measure,[56] and the Morisky Medication Adherence Scale (8 item)[57] constitute several examples of PROs assessing health-promoting behaviors.

Patient Experience of Care

Patient ratings of health care are an integral component of patient-centered care. In its definition of the essential dimensions of patient-centered care, the Institute of Medicine includes shared decision-making among clinicians, patients and families; self-efficacy and self-management skills for patients; and the patient's experience of care.[58,59] Measurement of patient ratings is a complex concept that is related to perceived needs, expectations of care, and experience of care.[60-67] Patient ratings can cover the spectrum of patient engagement, from experience to shared decision-making to self-management to full activation. Clinicians' recognition of patient preferences and values can help health care professionals to tailor treatments based on informed decisions that their patients might have. In fact, improving decision quality is one of the most important things that the nation can do to improve the quality (processes and outcomes) of health care and thus enhance value for health care expenditures. Thus, patient ratings of their experiences with care provide information very salient to patients and families, but they also have considerable policy implications. Each safe practice in the updated NQF consensus report includes a section titled "Opportunities for Patient and Family Involvement."[68]

The three major types of patient health care ratings relate to evaluations of patient satisfaction, patient motivation and activation, and patient reports of their actual experiences. Patient satisfaction is a multidimensional construct that includes patient concerns about the disease and its treatment, issues of treatment affordability and financial burden for the patient, communication with health care providers, access to services, satisfaction with treatment explanations, and confidence in the physician.[69 70-75] Shikiar and Rentz[65] proposed a three-level hierarchy of satisfaction: (1) satisfaction with health care delivery, including issues of accessibility, clinician-patient communication, quality of facilities; (2) satisfaction with the treatment regimen, including medication, dietary and exercise recommendations, etc.; and (3) satisfaction with the medication itself, rather than the broader treatment. Patient satisfaction has important implications for clinical decision-making and enhancing the delivery of health care services; it is increasingly the focus of research and evaluation of medical treatments, services and interventions.[76] It is an important indicator of future adherence to treatment.[64,77-82] Satisfaction has a long history of measurement and there are numerous available instruments.[62,67,83-91]

One potentially important predictor of health outcomes is patient activation, or the degree to which patients are motivated and have the relevant knowledge, skills, and confidence to make

optimal health care decisions.[92,93] Hibbard and colleagues have developed a 13-item scale, the Patient Activation Measure (PAM),[94,95] which has demonstrated favorable psychometric properties in several cross-sectional and some longitudinal studies.[93] While there has been increasing appreciation of the benefits of activated patients,[96] there is not yet commensurate support to help patients become more activated with respect to their healthcare decision-making.[95] While there is good support for the claim that improvements in patient activation are associated with improvements in self-reported health behaviors,[93,96] additional research is necessary to better understand these relationships and relevance to actual behavior. Patient activation, as measured by the PAM or otherwise, may be useful moderators or mediators of patient-reported outcomes that may be useful for performance measurement.

The newer focus is on measuring patient reports of their actual experiences with health care services.[97] Reports about care are often regarded as more specific, actionable, understandable, and objective than general ratings alone.[98,99] The Consumer Assessment of Healthcare Providers and Systems (CAHPS®) program is a multi-year initiative of the Agency for Healthcare Research and Quality (AHRQ) to support and promote the assessment of consumers' experiences with health care (www.cahps.ahrq.gov/About-CAHPS/CAHPS-Program.aspx) The CAHPS® program has two main goals: (1) to develop standardized patient questionnaires, and 2) to generate tools and resources that produce understandable and usable comparative information for both consumers and health care providers. The CAHPS project has become a leading mechanism for the measurement of patient perspectives on health care access and quality.[97]

III. Method and Mode of Administration, Data Collection and Analysis Issues

To accommodate the needs of patients with diverse linguistic, cultural, educational and functional skills, clinicians and researchers require some flexibility in choosing appropriate methods and modes of questionnaire administration for PROs.[100] Numerous issues complicated the scoring and analyzing of PRO response data. We first describe these methodological issues (see Table 3) and then discuss barriers.

Table 3. Main Characteristics of PRO Methods Issues[101,102]

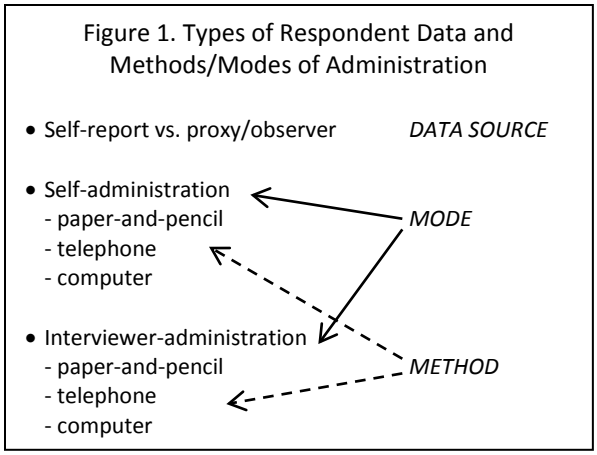
Methodological Issue	Main Characteristics	Strengths	Limitations
Source of report			
Self	• Person responds about him/herself	• Expert on own experience	• Not always possible to assess directly e.g., because of cognitive or communication deficits or age /developmental level

Methodological Issue	Main Characteristics	Strengths	Limitations
Proxy	• Person responds about someone else	• Useful when target of assessment unable to respond • Can provide complementary information	• May not accurately represent subjective or other experiences
Mode of administration			
Self	• Person self-administers PRO and records the responses	• Cost-effective • May yield more participant disclosure • Proceed at one's own pace	• Potential for missing data • Simple survey design (e.g., minimal skip patterns)
Interviewer	• Interviewer reads questions out loud and records the responses	• More complex survey design (e.g., skip patterns) • Useful for respondents with reading, writing or vision difficulties	• Interviewer costs • Potential for bias (interviewer bias, social desirability bias, acquiescent response sets)
Method of administration			
Paper-and-pencil	• Patients self-administer PRO using a paper and writing utensil	• Cost-effective	• Prone to data entry errors • Data entry, scoring requires more time • Less amenable to incorporation within EHR
Electronic	• Patient self-administers PRO using computer- or telephone-based platform	• Interactive • Practical • Increased comfort for socially undesirable behaviors • Minimizes data entry errors • Immediate scoring,	• Cost • Potential discomfort with technology • Accessibility • Measurement equivalence

Methodological Issue	Main Characteristics	Strengths	Limitations
		feedback • Amenable to incorporation within EHR	
Setting of administration			
Clinic	• Patients complete PROs when they arrive to clinic appointments	• Real-time assessment of outcomes • Feasibility with use of electronic methods of administration	• Impact on clinic flow • Interruptions resulting in missing data • Patient anxiety • Staff burden
Home	• Patients complete PROs at home prior to, or in between clinic visits	• Minimizes impact on clinic flow • Minimizes staff burden	• Accessibility • Health information privacy • Data security • Patient safety
Other	• Patients complete PROs at other types of settings (e.g., skilled nursing, rehabilitation)	• Feasibility with electronic methods of administration	• Cognitive capacity and potential need for proxy
Scoring			
Classical test theory	• Raw scores	• Easy to implement and understand	• All items must be administered
Modern test theory	• Probabilistic approach	• Enables CAT (tailored questions) • Shorter questionnaires with more precision	• Difficult to implement and understand

Methodological issues

Administering PRO instruments requires users to make decisions about three aspects of data collection: (1) the source of the information, (2) the recorder of the information (mode), and (3) the method used to capture the information (see Figure 1). Each of these actions is described below. These three aspects can also be combined in various ways, e.g., a patient might use the telephone to self-administer a PRO instrument, or an interviewer might use a computer to read questions and record answers.



Source: Self versus Proxy

The patient's perspective is the focal point of PRO assessment. In some circumstances, directly obtaining this perspective may be difficult or impossible. In adults, cognitive and communications deficits and burden of disease, for example, can limit potential subjects' ability to complete PRO questionnaires.[103] This is especially likely to occur with the elderly and with people of any age with severe disease or suffering from neurological disorders. Children's participation can be limited by these same factors plus issues specific to their age and developmental level.[103-105] Failing to include these populations can result in potentially misleading interpretations of results. Thus, attempting to include them in PRO assessment efforts is crucial; using all possible mechanisms for obtaining self-reports is a high priority, but accomplishing this may be out of the question for some populations.

One way to include the greatest number of patients is to use proxy respondents to obtain PRO information for patients who are unable to respond. Using either significant others (e.g., parents, spouses or other family members, friends) or formal caregivers (physicians, nurses, aides, teachers) as proxies can provide many potential benefits. They not only allow inclusion of a broader and more representative range of patients in the entire measurement effort, but they also can help minimize missing data and increase the feasibility of longitudinal assessment. The usefulness of proxy responses as <u>substitutes</u> for patient responses depends on the validity and reliability of proxy responses compared with those attributes for patient responses. When evaluating the quality of proxy responses, proxy responses are usually compared with patient responses. This is a reasonable approach, when proxy responses are being used to replace patient responses. Agreement between proxies and patient pairs is typically assessed at either the subscale level via the Intraclass Correlation Coefficient (ICC) or the item level by the kappa statistic, although other types of analyses have been advocated.[106] Patient and proxy responses are also often compared at the group level by comparing mean scores. Group comparisons help detect the magnitude and direction of any systematic bias that might be present.

Both the adult and pediatric literature suggests that agreement between proxy and patient ratings is higher when rating observable functioning or HRQL dimensions such as physical and instrumental activities of daily living, physical health and motor function; agreement is typically lower for more subjective dimensions such as social functioning, pain, cognitive status or function and psychological or and emotional well-being.[105,107-111] Using continuous rather than dichotomous ratings improves agreement.[112] Extent of disagreement increases with increasing age of adolescents,[113] and as the severity of patient illness, cognitive impairment or disability

rises[114-117] Type of proxy (e.g., parent versus caregiver), and proxy characteristics such as age, education, and level of stress may also affect agreement.[118,119] In terms of direction of disagreement, proxies for adults tend to rate them as having more symptoms, functional difficulties, emotional distress and negative quality of life; the main exception is pain, about which proxies tend to under-report.[107] Patterns of disagreement for child- versus proxy-reported outcomes are not consistent.[120] Even when self- and proxy-reports disagree for either children or adults, differences tend to be small.[120,121]

Proxy assessment may substitute for patient assessment where needed, but it may also complement it. Proxies can be asked to assess the patient as they think the patient would respond (i.e. proxy-patient perspective), or they can be asked to provide their own perspective on the patient's functioning or HRQL. This type of "additional" rating may be better described as either external- or other-ratings[122] for the sake of clarity. It is important that the measure makes clear which perspective is desired.[120] The external (i.e. "other") perspective may provide particularly relevant information when the person is unable to provide any self-assessment, but it can be important even when the patient can give his or her own answers. In such cases, patient-other agreement may not necessarily be desirable. For example, patients in the earlier stages of dementia may be able to provide responses to PROs but fail to recognize the extent of their impaired well-being and physical role functioning. In such cases, next-of-kin caregivers such as a spouse could provide a different ("external") assessment that indicates that the patient has a little or a considerable amount of problems in various ways, such as getting the groceries from car to kitchen, or being comfortable in a social setting. In these circumstances, external (proxy) respondents clearly can introduce clinically important information.

Mode: Self-administration versus Interviewer-administration

Self-administration of PRO questionnaires is neither expensive nor influenced by interviewer effects, for these reasons, this mode of administration has traditionally been preferred. However, self-administration is not feasible for some patient populations, such as those who may be too ill to self-administer a questionnaire. In these cases, interviewer-administration is often required. Until recently, interviewer-administration was also required for those with low literacy; however, new multimedia methods are now available to overcome this issue (see below).

Advantages and disadvantages of different modes of administration were summarized by Fowler[101] and Naughton[102] (see Table 3). Self-administered instruments are more cost-effective from a staffing perspective, and they may yield more patient disclosure, especially when collecting sensitive information.[123] Disadvantages include the potential for more missing data, and the inability to clarify any misunderstandings in questions or response options. By contrast, interviewer-administered instruments allow for probes and clarification, and they permit more complexity in survey design (e.g., the use of complicated skip patterns or open-ended questions). This mode is also useful for persons with reading, writing or vision difficulties. Disadvantages include the costs required to hire, train and supervise interviewers, and the potential pressure on respondents to answer quickly, rather than letting them proceed at their own pace. The potential for interviewer bias cannot be overlooked; it may arise from systematic differences from interviewer to interviewer or, occasionally, systematic errors on the part of many or even all interviewers.[124]

Other sources of bias for both administration modes include social desirability response set(the tendency to give a favorable picture of oneself) and acquiescent response set (the tendency to agree/disagree with statements regardless of their content).[125,126]

Legitimate concerns arise about the potential biasing effects of mode of administration on data quality and interpretation.[127] Overall, evidence supports high reliability for instruments administered with different modes, but response effects have varied and have not been consistently in the same direction.[121-124] For example, some studies have reported more favorable reports of well-being on self-administered questionnaires,[128] whereas others have found the opposite effect.[129-131] Still other studies reported mixed results[132] or found no important differences attributable to mode of administration after adjusting for other factors.[101,133,134] Fortunately, many types of error and bias can be overcome by appropriate selection and training of interviewers. Effects of different modes can also be evaluated with various psychometric and statistical techniques and models to determine the potential impact of response effects.[135-139]

Method of Administration

Advances in technology have changed the face of PRO assessment, increasing the number of administration options available. Multiple methods of self-report administration currently exist, and the different methods may have different effects on the quality of the data.[127] Although the different administration methods provide more options for researchers and clinicians, the different methods of administration require different skills and resources of people being asked to respond to the questionnaire; this means that the choice of method of administration may result in differing levels of respondent burden.[127] Several factors may account for differences in data quality across methods of administration: impersonality of the method, cognitive burden on the respondent, ability to establish the legitimacy of the reasons for which patients or others are even being asked to complete a questionnaire, control over the questionnaires, and communication style.[127] Thus, when users are deciding on one (or more) appropriate methods of administration for a given PRO, they must give these factors due consideration.

Historically, paper-and-pencil administration served as the primary method of PRO assessment. Many PROs were originally developed with the intention of paper-based administration, but they may be (and typically are) amenable to an electronic-based administration.[140] Paper-and-pencil remains a widely used PRO administration method, with its primary advantage being cost-effectiveness in situations in which there are few mailing and follow-up costs. However, the paper-and-pencil method has disadvantages. For example, it may require that a person's responses be manually entered into a database for scoring purposes, raising the possibility of data entry errors that threaten the integrity of the results. Similarly, the need for manual data entry and scoring can be time-intensive. Although the availability of optical mark recognition and optical character recognition allow for scanning of paper-and-pencil PROs, this process still requires an extra step on the part of staff and may limit the acceptability of paper-and-pencil administration for purposes in which timely scoring and interpretation is important.

Advances in technology and the increasingly widespread availability of electronic resources have provided several alternatives to the paper-and-pencil administration method. Advances in telephone technology have enabled the use of interactive voice response (IVR) to administer PROs. IVR involves a computer audio recording of PRO questions administered via telephone to which people indicate their responses by selecting the appropriate key.[127,140] In addition, computer-based administration methods have emerged as feasible alternatives to paper-and-pencil, such as web-based platforms, touchscreen computers, and multimedia platforms that can accommodate people with a range of literacy and computer skills (e.g., Talking Touchscreen/Pantalla Parlanchina, audiovisual computer-assisted self-interviewing[127,140-142]).

Newer mobile forms of technology such as tablet computers and smartphones also offer promise as newer generation methods of PRO administration.

The electronic administration methods have advantages that contribute to their increasingly widespread adoption. For example, because patients or respondents enter the data themselves, the opportunity for data entry errors is minimal compared with paper-and-pencil administration with separate data entry. These electronic methods also typically allow for immediate scoring and feedback, which enhances applications requiring timely results. Furthermore, electronic PRO administration has been shown to be practical, acceptable, and cost-effective.[52] Electronic methods may also provide people with increased comfort when responding to questions about socially undesirable behaviors.[143]

Nonetheless, these advantages must be considered in light of several important disadvantages. First, the cost of purchasing technology-based platforms may exceed that of traditional paper-and-pencil methods. Additionally, some patients may experience discomfort with technology or lack the skills necessary to navigate electronic administration methods. Moreover, reliance upon methods such as web-based platforms or smartphones raises questions about people's access to these technologies, if they are not provided in the relevant settings as part of clinical practice, quality improvement, or other assessment efforts.

The availability of multiple methods of PRO administration highlights the importance of measurement equivalence across methods.[140] Measurement equivalence is determined by comparing the psychometric properties of the data obtained via paper-based administration and the data collected through electronic-based administration.[140] It can be assessed via cognitive testing, usability testing, equivalence testing, or psychometric testing (or various combinations of these techniques).[140] A growing body of research documents the equivalence of electronic and paper-and-pencil administration of PROs.[144-146] These findings support the viability of electronic PRO administration as an alternative to paper-and-pencil methods.

In addition to measurement equivalence, patient privacy is another concern that cuts across both paper-and-pencil and electronic administration methods, albeit in differing ways. For paper-based PROs, physical transfer of the PRO measure from patient to provider, as well as the physical existence of the completed PRO instruments may pose risks to the privacy and confidentiality of patients' responses. Privacy also emerges as a concern about electronic-based methods, given potential security breaches related to transfer of data, computer errors, or unauthorized access to patient-reported data. These threats underscore the need for reliable and secure electronic platforms to protect patients' privacy in the context of PRO assessment.

PROs in the Clinical Setting

Collection of PRO data as part of clinical care has become more common.[147-150] There are many benefits of facilitating the introduction of these PROs into clinical practice and decision-making. Advocates for the use of PROs in clinical care propose that the results assist clinical providers management of patients' care,[151] enhance the efficiency of clinical practice,[146,152] improve patient-provider communication,[146,152-154] identify patient needs in a timely manner,[146,155] and facilitate patient-centered care.[146] However, other findings suggest regional variation in perceived health and no positive effect of feedback via PROs on care, even when combined with guideline-recited interventions.[156-158] As PROs are used more in clinical practice, some methodological issues pertaining to the settings in which they are administered merit consideration.

A growing number of studies have investigated the use of PROs in the clinic setting.[146,152,154,155,159-162] When selecting PROs for administration in clinical practice, users need to consider the efficiency of PRO administration, scoring, and interpretation. These factors are especially important because of the time-sensitive nature of the clinic workflow.[146,159] In addition, acceptability of both the PRO measures and the data collection process for both patients and practices' staff is essential.[146,159,163]

Historically, several barriers have impeded the widespread implementation of PRO data collection in clinical settings of all sorts, but especially smaller or private practices. Many drawbacks are associated with paper-and-pencil administration of PROs. One such barrier involves concerns about the potential disruption to the clinical work flow if patients are asked to complete PROs.[151] For example, collection of PRO data in clinical settings may be impeded by staff burden and clinician disengagement.[151] Fortunately, technology advances, and the increased opportunities for methods of PRO administration that they afford, may help to overcome some barriers to PRO data collection in clinical practices and settings.[159] For example, research supports the feasibility of using tablet computers[146,155] and touchscreen computers for these purposes.[141,142,154,159,160] Use of computers to administer PROs may streamline and expedite the process, and minimize staff burden and impact on clinic flow. Conversely, concerns arise regarding the impact of clinic flow on the integrity of data collection, given the potential for patients to be interrupted while completing PROs, which could potentially result in missing data.[151] Another potential barrier involves the possibility that patients may experience anxiety in completing PRO measures in the clinical settings before their appointments.[151] Similarly, a possible lack of privacy when completing PROs in waiting rooms or similar circumstances poses another potential obstacle to adequate PRO administration. Many of these concerns can be addressed by incorporating PROs as they apply to PRO-PMs into the clinical work flow. This will also enhance completion rates. Both patients and providers will then be more likely to see this effort as integral to patient care.

Completing PRO measures from home before or between medical appointments has been proposed as one strategy to overcoming the problems outlined above. [151,164,165] Both web-based PRO administration and IVR constitute possible methods for at-home PRO data collection.[151,161,162] Although the home may serve as a feasible alternative to the clinical practice setting for various reasons, those considering implementing home-based PRO data collection need to consider several factors.[151,164] First, for patients to be able to complete PRO measures at home, they must have access to the type of technology by which the PRO is administered (e.g., internet). Second, patients must find completing PRO measures at home acceptable. Users should have a plan in place to address situations in which home-based PRO responses suggest critical or acute problems. This may pose a logistical challenge in comparison with PROs completed in-clinic, where medical providers and access to intervention are readily available.

As with any setting, health information privacy is paramount; therefore, one barrier to home-based PROs is availability of secure data collection platforms.[151 166] Finally, an especially difficult issue may be clinician acceptability of home-based PRO data collection. The problems include reimbursement for clinicians' time using a website to address patient-reported outcomes, rather than meeting directly with patients to discuss questions or problems that their patients raise through answers to the PRO instruments.[151,166]

Implementing PRO data collection in other settings, such as rehabilitation or skilled nursing facilities, may also yield valuable clinical information and guide interventions. Less research has addressed the methods issues in administering PROs in these settings. However, handheld

technology has been proposed as a means of facilitating collection of PRO data in the rehabilitation setting following orthopedic surgery.[167]

Apart from technology per se, other issues in such facilities include the varying level of patients' acuity status and levels of cognitive capacity to complete PRO instruments. In these cases, users may need to consider whether using proxy reports may be beneficial. In any case, the potential strengths and weaknesses of different modes and methods of administration still need to be taken into account.

Scoring: Classical Test Theory versus Modern Test Theory

PROs are "latent (not directly observable) variables." The only way to estimate a person's level on a particular attribute is by asking questions that are representative of that attribute. Most PRO instruments comprise multiple items that are aggregated in some way to produce an overall score. The most common multi-item instruments are designed to reflect a single underlying construct. The item responses are either caused by or are manifestations of the underlying latent attribute, and the items are expected to correlate with one another.[168-171] There are other kinds of multi-item measures in which the items may cluster together, but would not be expected to correlate. A common example of this latter measure is a "comorbidity index" comprised of various health conditions, e.g., diabetes, asthma, heart disease, etc. Another example might be the creation of a measure of access to care consisting of problems with paying for care, having a regular provider, ease of transportation to care and ease of making an appointment. Although such items would not necessarily be correlated, together they might form an adequate measure of access. The discussion on scoring below refers to the former type of instrument.

Scoring is based on classical test theory (raw scores) or modern test theory (item response theory; IRT).[172-181] Multiple items are preferred because a response to a single item provides only limited information to distinguish among individuals.[182] In addition, measurement error (the difference between the "true" score and the "observed" score) tends to "average out" when responses to individual items are summed to obtain a total score.[182-184]

Classical test theory estimates the level of an attribute as the sum, perhaps weighted, of responses to individual items, i.e., as a linear combination.[10,182,185-188] This approach requires all of the items on a particular PRO instrument to be used in every situation for it to be considered valid, i.e., the instrument is "test-dependent"[186,188-190] IRT, by contrast, enables "test-free" measurement, i.e. the latent trait can be estimated using different items as long as their locations (difficulty levels) have been calibrated on the same scale as the patients' ability levels.[10,182,188-191 182,192,193] IRT allows computer-adaptive testing (CAT) in which the number, order and content of the questions are tailored to the individual patient. This approach has two distinct advantages: (1) questionnaires can be shorter, and (2) the scale scores can be estimated more precisely for any given test length. This also means that different patients do not need to complete the same set of items in every situation.[10]

Using IRT poses nontrivial challenges, however, Understanding the assumptions and the psychometric jargon - e.g., "calibration," "difficulty levels" - is not easy. The methodology and software are complex. IRT is also not appropriate for causal variables and complex latent traits.[10,188,189,194] Overall, however, IRT offers a very convenient and efficient framework for PRO measurement and it is becoming increasingly well understood and easier to adopt.

Linking or Cross-talk Between Different Measures of the Same Construct

A common problem when using an array of health-related outcomes for diverse patient populations and subgroups is establishing the comparability of scales or units on which the outcomes are reported.[195,196] The emphasis has typically been focused on the metric over the measure. “Equating” is a technique to convert the system of units of one measure to that of another. This process of deriving equivalent scores has been used successfully in educational testing to compare test scores obtained from parallel or alternate forms that measure the same characteristic with or without having common anchor items.

Theoretically (and in practice when certain conditions are met), different age-specific measures could be linked, thus placing child, adult, and geriatric estimates on a common metric. For example, the many items that constitute a condition-specific (e.g., cancer) quality of life scale could be incorporated into a single shared bank and linked through a common-anchor design.[195] The methods of establishing comparable scores (often called “linking”) vary substantially depending on the definition of comparability. For that reason, standardization is critical in comparing PROs across studies. Two measures may be considered linked if they produce scores that match the first two moments (i.e., mean and SD) of their distributions for a specific group of examinees or two randomly equivalent groups. Another definition may involve matching scores with equal percentile ranks based on a single sample of examinees or random samples drawn from the same population.

Addressing Barriers to PRO Measurement

Yet other barriers to PRO measurement need to be addressed. These include: administering PROs in vulnerable populations; literacy, health literacy and numeracy; language and cultural differences; differences in functional abilities; response shift; use of different methods and modes of administration; and the impact of non-responders to items and questionnaires. In the review below, we also note best practices and recommendations for addressing these barriers.

Vulnerable Populations

Recognition is growing that some population subgroups are particularly vulnerable to receiving suboptimal health care and to achieving health outcomes equivalent to those experienced by the general population. [197-199] Vulnerability is multifaceted. It can arise from age, race, ethnicity, or sex (or gender); health, functional, or developmental status; financial circumstances (income, health insurance); place of residence; or ability to communicate effectively.[197] Moreover, many of these factors are synergistic, so that vulnerability has many sources that present a complicated picture for persons in these groups. This definition encompasses populations who are vulnerable because of a chronic or terminal illness or disability and those with literacy or language difficulties.[141,198] It also includes people residing in areas with health professional shortages.[168]

Administration of PRO questionnaires is usually performed with paper-and-pencil instruments, and multilingual versions of questionnaires are often not available. Interviewer administration is labor intensive and cost prohibitive in most health care settings. Therefore, patients with low literacy, those with certain functional limitations, or those who do not speak English are typically excluded, either explicitly or implicitly, from any outcome evaluation in a clinical practice setting in which patient-reported data are collected on forms.

As PROs continue to play a greater role in medical decision making and evaluation of the quality of health care, sensitive and efficient methods of measuring those outcomes among underserved populations must be developed and validated. Minority status, language preference, and literacy level may be critical variables in differentiating those who receive and respond well to treatment from those who do not. These patients may experience different health outcomes because of disparities in care or barriers to care. Outcome measurement in these patients may provide new insight into disease or treatment problems that may have gone undetected simply because many studies have not been able to accommodate the special needs of such patients.[198,200]

Literacy

Low literacy is a widespread but neglected problem in the United States. The 1992 National Adult Literacy Survey (NALS)[201] and the 2003 National Assessment of Adult Literacy (NAAL)[202] measured three kinds of English language literacy tasks that adults encounter in daily life (prose literacy, document literacy, quantitative literacy). Almost half of the adult population experiences difficulty in using reading, speaking, writing, and computational skills in everyday life situations. An additional seven million adults in the U.S. population were estimated to be non-literate in English. "Health literacy," is "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions." [203] This involves using a range of skills (e.g., reading, listening, speaking, writing, numeracy) to function effectively in the health care environment and act appropriately on health care information. [204,205] Limited health literacy is widespread,[204,206] and is associated with medication errors, increased health care costs, hospitalizations, increased mortality, decreased self-efficacy, and inadequate knowledge and self-care for chronic health conditions.[204,207-210] Health literacy may be more limited than functional literacy because of the unfamiliar context and vocabulary of the health care system.[204,211]

Contributing to poor understanding of the importance of literacy skills is the fact that low literacy is often underreported. The NALS reported that 66% to 75% of adults in the lowest reading level and 93% to 97% in the second-lowest reading level described themselves as being able to read or write English "well" or "very well."[201] In addition, low-literacy individuals are ashamed of their reading difficulties and try to hide the problem, even from their families.[212,213] Lack of recognition and denial of reading problems creates a barrier to health care. Because they are ashamed of their reading difficulties, low-literacy patients have acknowledged avoiding medical care.[212,213] In addition, because everyday life may place only moderate reading demands on people, individuals may not even be aware of their reading problems until a literacy-challenging event occurs (e.g., reviewing treatment options, reading a consent document, completing health assessment forms).[212,213]

A reader's comprehension of text depends on the purpose for reading, the ability of the reader, and the text that is being read. Two important factors in the readability of text are word familiarity and sentence length.[214] Unfamiliar words are difficult when first encountered. Long sentences are likely to contain more phrases or clauses. Although longer sentences may communicate more information and more ideas, they are more difficult for readers to manage than more, but shorter, sentences that convey the same information. Moreover, longer sentences may also require the reader to retain more information in short-term memory.[215-218]

Addressing health literacy is now recognized as critical to delivering person-centered health care.[219] It is an important component of providing quality health care to diverse populations, and will be incorporated into the National Standards for Culturally and Linguistically Appropriate

Services.[220] For example, it is challenging to translate highly technical medical and legal language into easily understood languages, whether it is English or another language. Health literacy practices are also included in the National Quality Forum 2010 updated set of safe practices.[68] A recent discussion paper summarized 10 attributes that exemplify a "health literate health care organization."[219] These attributes cover practical strategies across all aspects of health care, from leadership planning and evaluation, to workforce training, to clear communication practices for patients.

Language and Culture

The availability of multiple language versions of PRO questionnaires has enabled users to administer them relatively routinely in diverse research and practice settings. For various purposes, doing analyses on data that have been pooled across all patients is desirable. Yet concern is often voiced about combining data from different cultures or languages.[2] In some research and practice-based initiatives, evaluating cross-cultural differences in PROs is of interest. In all these applications, using unbiased questionnaires that can detect important differences among patients is critical.[198,221,222]

Possible cultural differences in interpreting questions and in response styles may limit data pooling or may constrain comparisons across members of different cultural groups.[223-225] Similarly, poor quality translations can produce non-comparable language versions of PRO questionnaires.[224,226] [224,227] The extent to which items in a questionnaire perform similarly across different groups (e.g., the extent to which they are cross-culturally or cross-linguistically equivalent) is of critical interest when determining whether the questionnaire can be used as an unbiased measure of a PRO. [222] [228-239] Without assurances that the PRO questionnaire is culturally and linguistically "fair," detected treatment differences caused by items that function differently across groups could incorrectly be interpreted to reflect real treatment differences. Similarly, differences in questionnaire performance may mask true treatment differences, especially when language or cultural groups are not balanced across the populations, practices or settings to be compared.

Functional Abilities

Ideally, PRO instruments that are intended to be used in performance measurement applications can be completed by all patients in the target populations. Otherwise, if a significant proportion of the population is left out, the remaining individuals being assessed may be unrepresentative of the whole practice or setting; this problem can (and probably will) compromise the validity of the performance measure.

Functional limitations associated with disability are one type of potential barrier to PRO assessment that could affect PRO use in performance measures. The prevalence of disability, defined as specific functional or sensory limitations, is estimated at 47.5 million Americans, or 22% of the U.S. population.[240] People with disability are more likely to develop health conditions and be consumers of health care than those with no disabilities of these types. Thus, they are an important group to include when evaluating health care, but one that is frequently not included in such clinical, quality improvement, or simulative initiatives.[241,242]

Common disabilities that can affect PRO assessment include problems with vision (e.g., decreased visual acuity, color-blindness), hearing, motor skills (e.g. upper extremity limitations), and cognitive deficits (e.g., impaired comprehension, reading). Fortunately, many of these barriers can be addressed by a variety of techniques: choosing appropriate methods and modes

of data collection, enabling use of assistive devices and technology, and using principles of "universal design" when developing instruments [201-202].

Universal design refers to designing products and environments in such a way as to be usable by all people, to the greatest extent possible, without adaptation or specialization.[243,244] A well-known example of universal design is the use of curb cuts. Initially intended to facilitate the use of wheelchairs, curb cuts have also benefited bicycle riders and children in strollers, among others. An exhaustive examination of how the principles of universal design can be applied to PRO assessment is beyond the scope of this paper, and those developing or modifying measures according to the principles of universal design are encouraged to consult with relevant experts. Also, if developers are creating an instrument based on information technologies, using the standards included in Section 508 of the Rehabilitation Act Amendments of 1998 can maximize flexibility.[245] Although we cannot list all potential ways to address functional limitations, we identify below some common ways to do so. Harniss and colleagues describe how PROMIS® is taking a systematic approach to enhancing accessibility.[246]

In general, providing multiple means of understanding and responding to measures is important; these include visual, voiced, and tactile mechanisms. The specific means may differ depending on the method and mode of administration. Thus, for people with impaired vision one might consider using in-person or telephone interviews (advantages and disadvantages discussed in an earlier section), an IVR system, Braille responses for Braille users, or touchscreen with tactile or audio cues. Information technology-based systems should accommodate assistive devices such as screen readers and screen-enlargement software. For patients with hearing impairments, options include providing visual presentation of words or images, using TTY or a Video Relay Service, and allowing the user to adjust the sound level. For persons with motor limitations, response modes that are easier to manipulate (track ball) or are non-motoric (e.g. using voice recognition software) can be helpful. For those with certain types of cognitive deficits (e.g., limited reading comprehension) the methods to address literacy described earlier should be considered. However, if cognitive deficits are severe, a proxy respondent may be more appropriate (also discussed above).

Allowing for multiple response modes or methods may lead to measurement error. In a later section, we discuss the potential impact of different methods and modes on response rate, reliability and validity. The risk of introducing measurement error seems outweighed by the risk of excluding a significant segment of the population.

Response Shift, Adaptation, and Other Challenges to Detecting True Change

The ability to detect true change over time in PROs poses another barrier to the integrity of PRO assessment. Often, detecting true change is associated with the phenomenon of response shift, which has been defined as, "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (i.e. scale recalibration); (b) a change in the respondent's values (i.e. the importance of component domains constituting the target construct) or (c) a redefinition of the target construct (i.e. reconceptualization)" (p.1532).[247] A change in perspective over time may result in patients' attending to PROs in a systematically different way from one time point to another.[248]

Response shift serves as a barrier to PRO assessment for several important reasons. For example, it threatens longitudinal PRO assessment validity, reliability, and responsiveness.[248-251] Response shift can complicate the interpretation of PRO outcomes, since a change in PRO outcome may occur because of response shift, an effect of treatment, or both.[252]

Monitoring for response shift can aid PRO users in interpreting longitudinal PRO data.[250] Several strategies have been proposed to identify response shift, although each has limitations. The "then test" compares an actual pre-test rating and a retrospective pre-test rating to assess for shift, but it is less robust than other methods of detecting response shift,[248] and is confounded with recall bias.[251] Structural equation modeling has also been proposed as a way to identify response shift; however, it is sensitive only if most of the sample is likely to make response shifts.[253] Finally, growth modeling creates a predictive growth curve model to investigate patterns in discrepancies between expected and observed scores, thus assessing response shift at the individual level.[254] Although growth modeling enables users to detect both the timing and shape of response shift,[250] it cannot differentiate between random error and response shift.[251]

Implications of the Different Methods and Modes on Response Rate, Reliability and Validity

Data Collection Methods

Decisions must be made related to the data collection method and the implications of those decisions on costs and errors in surveys.[123] Two basic issues underlie these decisions: (1) What is the most appropriate method to choose for a particular question? and (2) What is the impact of a particular method on survey errors and costs?

Different methods differ along a variety of dimensions.[123] This includes, although are not limited to, the degree of interviewer involvement and the level of interaction with the respondent. Channels of communication (sight, sound, touch) used can be critical; various combinations may prompt different issues of comprehension, memory stimulation, social influence affecting judgment, and response hurdles). Finally, the degree of technology use is a major consideration.

Using a Different Method or Mode than Originally Validated

Considering the implications of using a different method or mode than the one on which the PRO was originally validated is also important. Many existing PROs were initially validated in paper-and-pencil form. However, potential differences exist between paper-and-pencil and electronic-based PRO administration,[144] ranging from differences in how items and responses are presented (e.g., items presented one at a time, size of text) to differences in participant comfort level in responding (e.g., ability to interact with electronic –based platform).[144]

As noted earlier, a growing body of research suggests measurement equivalence between paper- and computer-administered PROs.[144,255] However, the effect of a particular data collection method on a particular source of error may depend on the specific combination of methods used.[123] Thus, as new methods are developed, studies comparing them with the methods they may replace must be done. Theory is important to frame expectations about the likely effect of a particular approach. Theory is informed by past mode-effects literature and by an understanding of the features or elements of a particular design.[123] Similarly, mode choices involve trade-offs and compromises. As such, the choice of a particular approach must be made within the context of the particular objectives of the survey and the resources available.[123]

Implications of Using Multiple Methods and Modes

The implications of using multiple methods and modes also warrant consideration. One might choose to blend methods for one or more reasons: cost reduction, faster data collection, and optimization of response rates.[123] When combining methods or modes (or both), users must ensure that they can disentangle any effects of the method or mode from other population characteristics. This is especially true when respondents choose which method or mode they prefer or when access issues determine the choice of method or mode.[123] As in the case of using a different method or mode than the one in which the PRO instrument was originally validated, instruments and procedures should be designed to ensure equivalence across both methods and modes.[256]

Accounting for the Impact of Non-responders

Difficulties with data collection and questionnaire completion are major barriers to the successful implementation of PRO assessment. The principal problem is that data that are missing can introduce bias in analyses, findings, and conclusions or recommendations.[10] The choice of mode and method of questionnaire administration can affect nonresponse rates and nonresponse bias.[123] In addition, often the timing of the assessment can be very important, e.g., just before or just after surgery.

Missing data may be classified as either item non-response (one or more missing items within a questionnaire), or unit non-response (the whole questionnaire is missing for a patient). Evaluating the amount, reasons and patterns of missing data is important.[257-260] Some common strategies to evaluate non-response bias include:

- Conducting an abbreviated follow-up survey with initial non-respondents[123]
- Comparing characteristics of respondents and non-respondents[261,262]
- Comparing respondent data with comparable information from other sources; [263] and
- Comparing on-time vs. late respondents.[264]

When dealing with missing data, analysts can use various statistical methods of adjustment. For item non-response in multi-item scales, several techniques are useful and tend to yield unbiased estimates of scores, e.g., simple mean imputation, regression imputation, and IRT models. For both item and unit non-response it is important to determine whether missing data are considered to be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR).[257,258] For unit non-response, there is a range of statistical techniques that could be implemented, depending on the reason for missing data.[265-269]

IV. Selection of Patient-Level PRO Measures

Patient-Centered Outcomes Research

An essential aspect of patient-centered outcomes research (PCOR) is the integration of patient perspectives and experiences with clinical and biological data collected from the patient to evaluate the safety and efficacy of an intervention (www.pcori.org). Such integration recognizes that although traditional clinical endpoints such as laboratory values or survival are still very important, we also need to look at how patients' health-related quality of life (HRQL) is affected by the disease and treatment. For such HRQL endpoints, in most cases, the patient is the best source for reporting what they are experiencing. The challenge is how best to capture patient data in a way that maximizes our ability to inform decision making in the research, healthcare delivery, and policy settings.

Access to psychometrically sound and decision-relevant PROs will allow clinicians, investigators, administrators and others to collect empirical evidence on the differential benefits and harms of a health-related intervention.[270-273] Those obtaining such information can then disseminate findings, as appropriate for the purpose, to patients, clinicians and health care professionals, payers or insurers, and policy makers to provide a richer perspective on the net impact of interventions on patients' lives using endpoints that are meaningful to the patients.[274]

Increasingly, longitudinal observational and experimental studies have included PRO measures. To optimize decision making in clinical care, these PROs must be assessed in a standardized way using questionnaires that demonstrate specific measurement properties.[270,273,275-278] Our group recently identified minimum standards for the design or selection of a PRO for use in PCOR activities.[279] Central to this work was an understanding of the critical attributes for which a PRO is judged to be appropriate or inappropriate for such purposes. We identified these standards through two complementary approaches. The first was to conduct an extensive review of the literature including both published and unpublished guidance documents. The second was to assemble a group of international experts in PROs and PCOR efforts to seek consensus on the minimum standards. [279]

Attributes of PRO Measures

Many documents summarize attributes of a good HRQL measure. They include (an illustrative list) guidance documents from the U.S. Food and Drug Administration (FDA); [280-283] the 2002 Medical Outcomes Trust guidelines on attributes of a good HRQL measure;[284] the extensive, international expert-driven recommendations from COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments);[276,285-289] the European Organization for Research and Treatment of Cancer (EORTC) guidelines for developing questionnaires; [290] the Functional Assessment of Chronic Illness Therapy (FACIT) approach;[28] the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force recommendation documents;[140,232,291,292] and several others.[236,275,293-295] The NIH PROMIS® network released a standards document in 2012 that is useful for informing the minimal and optimal standards for designing PRO measures.[296] In addition, ISOQOL recently completed two guidance documents on use of PROs in comparative effectiveness research and on integrating PROs in healthcare delivery settings that were relevant for this landscape review.[275,297]

Table 4, at the end of this section, presents long-established criteria or characteristics to consider in selecting PROs. It also notes some best practices for evaluating PROs for use in performance measurement.

Selecting PROs for use in performance measurement and related activities such as quality improvement programs raises the question of what are the key differences, if any, when selecting PROs for research purposes rather than these other non-research purposes. Generally speaking, the factors to consider when selecting PROs for these two kinds of activities are more similar than different. Thus, we focus here more on the differences that users will need to take into account.

One key difference involves the length of the PRO instrument. Longer PRO tools with more items may be better tolerated in the context of research than in clinical practice settings; thus, the feasibility and acceptability of using PROs for performance measurement demands shorter instrument length to facilitate widespread adoption. Addressing the need for shorter PRO

measures may, however, compromise other important measurement characteristics, such as precision.

Another key difference in factors to consider when selecting PROs for clinical practice quality improvement, or performance measurement and accountability efforts is the implications or consequences of the PRO data. Specifically, using PROs for these purposes carries the expectation that important consequences will arise in terms of accountability for health care professionals, health care systems and plans, and clinical settings. Therefore, the stakes of PROs are high in the performance measurement context, and they are higher than for research applications. The problem lies, in part, in the constraints to the quality of the measurement level arising from factors unique to performance measurement, such as length, or representativeness of the patient or consumer populations surveyed. These considerations highlight the importance of emphasizing responsiveness and sensitivity to change when considering PROs for use in the ways envisioned for NQF-endorsed measures.

In selecting a PRO for these practice, quality of care, and accountability purposes, a logical first step involves reviewing what measures already have been used successfully. Using PROs for these programs remains an under-studied area, but several examples of PROs used as indexes of performance measurement provide an initial foundation upon which the field can expand. The Veterans Health Study may be the best illustration of this point. It was developed to assess PROs within the VA system.[298] In response to the Veterans Health Administration's incorporation of patient-reported functional status as a domain of interest in their performance measurement system, the Veterans RAND 36 Item Health Survey (VR-36) and the Veterans Rand 12 Item Health Survey (VR-12) have been administered within the VA system to evaluate veterans' needs and to assess outcomes of clinical care at the hospital, regional, and healthcare system levels.[298,299] The Centers for Medicare & Medicaid Services (CMS) and its Medicare Advantage Program[300] have applied these methods for similar purposes, and CMS has also designated the VR-12 as the principal outcomes measure of the Medicare Health Outcomes Survey (HOS).[301]

Research examining the VR-36 and SF-36 in such uses does inform the selection of PROs for performance measurement, but limitations remain to the use of these measures as indicators of high-quality care and as sources of information for holding practices, providers, hospitals, health plans, or others accountable for their results. These limitations include their "static" nature which requires all items to be administered for analysts to be able to obtain an individual's score; this is true, even if some items add little to the precision of measurement. In addition, content is fixed by the composition of the scale.

Therefore, attention has turned to alternative PRO tools with clear potential for these types of uses (i.e., as patient-reported performance measures). PROMIS® constitutes arguably the best example of a future direction of PROs that will be acceptable for use in practice, quality improvement, or performance measurement programs. Developed using IRT methodology, PROMIS® offers a new generation of PRO measures with better reliability, validity, precision, and other attributes than is typically true for so-called legacy instruments; these measures have the important attribute of being shorter than such older instruments as well.[179] PROMIS® PRO measures form a hybrid between static generic PROs and more flexible adaptive measures that comprise items specific to measure content, but that are applicable across the diverse spectrum of health status. Although a growing body of literature provides preliminary evidence supporting the psychometric adequacy of the PROMIS® measures, future work is needed to explore the application of PROMIS® measures as performance measure PROs. Nevertheless, PROMIS®

system provides a model by which the use of PROs as performance measures can be expanded and elaborated upon, owing to its rigorous methodological characteristics.

Documentation, in peer-reviewed literature or on publicly accessible websites (or both), of the evidence of a PRO to reflect these measurement properties will result in greater acceptance of the PRO for use as performance measures. To the extent that the evidence came from populations similar to the studies' target population, the more confidence clinicians, analysts, administrators, or policymakers can have in the PRO to capture patients' experiences and perspectives.

Applying any set of selection standards for PROs calls for attention to several considerations. One key issue is that the populations involved in these efforts will likely be quite heterogeneous. This population heterogeneity should be reflected in the samples that participate in the evaluation of the measurement properties for the PRO. For example, both qualitative and quantitative studies may require quota sampling based on race and ethnicity that reflects the prevalence of the condition in the study target population. Additionally, patients must be actively engaged as stakeholders in the identification of the domains most important to measure via PROs, as well as in the selection of PRO measures for use in performance measurement.

Literacy demand is also an important consideration for use of PROs. Data collected from PRO measures is only valid if the participants in a study can understand what is asked of them and can provide a response that accurately reflects their experiences or perspectives. It is critical that developers of PRO measures be attentive to make sure the questions and response options are clear and easy to understand. Pre-testing of the instrument (e.g., cognitive testing) should include individuals with low literacy to evaluate the questions.[302]

Response burden must be considered when selecting a PRO measure and using it in a PCOR study. The instrument must not be overly burdensome for patients as they are often sick and cannot be subjected to long questionnaires or be asked to provide repeated, longitudinal data that may significantly disrupt their lives.

Finally, researchers must carefully consider the strength of evidence for the measurement properties. There is no threshold for which an instrument is valid or not valid for any or all populations or applications. In addition, there can be no single study that confirms all the measurement properties for all contexts. Like any scientific discipline, measurement science relies on an iterative, accumulating body of evidence examining key properties in different contexts. Thus, it is the weight of the evidence that informs the evaluation of the appropriateness of a PRO. Older PROs will have the benefit of having more evidence than more recent PROs; yet the newer PROs tend to have improved basic measurement properties that warrant attention.

PRO Characteristics for Consideration

Generic versus Condition-specific Measures

One primary factor to consider when selecting a patient-level PRO measure is whether to use a generic versus a condition-specific PRO. Several elements inform the selection of measures.[303] First, the specific population of interest may guide whether one opts to use a generic or condition-specific PRO. For example, if the target population comprises mainly healthy individuals, a generic measure may be the preferred choice. Conversely, if the goal is to

examine a specific subset of patients with a particular health concern, then a condition-specific measure may be more appropriate.

Second, outcomes of interest may guide the selection process. Generic measures may capture a different category of outcomes when compared to a condition-specific PRO. For example, a generic measure may assess domains of general quality of life, whereas a condition-specific PRO may measure symptoms expected to be directly addressed by a condition-specific intervention.

Third, the assessment purpose will influence the selection of generic versus specific measures. An excellent example of this stems from FDA guidance which states that pharmaceutical company claims of improved QOL must be specific to the QOL domain that was measured; the agency recommends that assessment of specific symptoms is an appropriate starting point.[304]

Generic PRO measures have several important advantages. They allow for comparability across patients and populations,[303] although they are more suitable for comparison across groups than for individual use.[305] Global PROs also allow assessments in terms of normative data which can be used to interpret scores.[303] This enables evaluation against population norms or comparison with information about various disease conditions. They can also be applied to individuals without specific health conditions, and they can differentiate groups on indexes of overall health and well-being.[303]

Generic PROs also have several disadvantages. They have tended to be less sensitive to change than condition-specific measures; for that reason, they may underestimate health changes in specific patient populations.[306] Additionally, they may fail to capture important condition-specific concerns[306] when applied in specific disease populations.

Condition-specific PROs are an alternative to generic PROs. One advantage of condition-specific PROs is greater sensitivity to change, because they focus on the concerns pertinent to the given condition.[303] They also enable differentiation of groups at the level of specific symptoms or patient concerns.[303] Almost by definition, however, the condition-specific focus introduces the notable difficulty of making comparisons across patient populations with different diseases or health conditions.[303]

Given their respective unique benefits and limitations, we conclude and recommend that a combination of generic and condition-specific measures is likely to be the best choice for the performance measurement purposes that NQF has most in mind. Generic and condition-specific PRO measures may measure different aspects of QOL when administered in combination,[307] resulting in more comprehensive assessment. Consequently, hybrid measurement systems have emerged to facilitate combining them. For example, the FACIT system consists of a generic HRQL measure plus condition-specific subscales. PROMIS®, which was developed to create item banks that are appropriate for use across common chronic disease conditions,[308] represents another example of a hybrid system of PROs that combine both global and targeted approaches.

Measurement Precision

Another factor to consider when selecting a patient-level PRO measure is measurement precision. Measurement precision refers to the level of variation in multiple measurements of the same factor; measures with greater precision vary less across assessment time points. PROs

with greater measurement precision also demonstrate greater sensitivity to change.[309] Given that most PROs were originally developed as research tools, they may lack the level of precision necessary for assessing individuals on these types of outcomes.[310] Although performance measures will aggregate to practice, provider or organization levels, adequate measurement precision at the patient level is still needed.

When considering measurement precision in selecting PROs, measures based on IRT tend to have greater precision than measures based on classical test theory.[310] Specifically, computerized adaptive tests (CATs) offer greater precision than static short-forms derived from item banks; however, short forms are an acceptable alternative when CAT approaches are infeasible.[311,312] Although CATs include a greater number of items in an item bank, they allow tailored measurement, resulting in shorter instruments and better precision. Consequently, using PROs derived from IRT techniques is recommended to achieve the greatest measurement precision.

Sensitivity to Change or Responsiveness

Sensitivity to change (also denoted responsiveness) constitutes another important factor to consider when selecting a PRO measure because the ability to detect a small, but important change is necessary when monitoring patients and implementing clinical interventions.[30] Sensitivity to change is a type of validity characterized by within-subject changes over time following an intervention.[313,314]

Responsiveness is conceptualized in many ways, which leads to different findings and interpretations.[315] Definitions of sensitivity to change range from the ability to detect any kind of change, regardless of meaningfulness (e.g., a statistically significant change post-treatment), to the ability to detect a clinically important change. To be clinically useful, PROs must demonstrate sensitivity to change both when individuals improve and when they deteriorate.[314]

Methods for assessing responsiveness vary markedly as well. These methods differ primarily in terms of whether they are intended to demonstrate statistically significant changes to quantify the magnitude of change.[315] The lack of equivalence across methods for detecting change can be problematic for interpretation, given that the different methods for detecting responsiveness produce different classifications of who is improved or not.[316] However, relying solely on statistical tests of responsiveness is not recommended, given that such findings may not accurately reflect what is meaningful to patients or clinician.[317]

Several factors can limit a PRO measure's sensitivity to change. First, multi-trait scales containing items that are not relevant to the population being assessed may fail to capture change over time.[318] The responsiveness of a PRO measure may also be constrained by using scales that offer categorical or a limited range of response options.[318] PRO measures that specify an extensive timeframe for reporting also will not be likely to demonstrate change, particularly when administered regularly over a brief period of time.[318] The responsiveness of a PRO measure is also limited when items that reflect stable characteristics are included, because these are unlikely to change. Scales that contain items with floor or ceiling effects are also problematic.[318] A PRO measure's sensitivity to change may depend upon the direction of the change. For example, Eurich and colleagues found that PROs were more responsive to change when patients got better clinically than when they got worse.[30]

In addition to these factors, a growing body of research suggests that condition-specific PROs are more sensitive to change than generic PROs.[30,32,319-321] This reflects the fact that

responsiveness to change is likely influenced by the purpose for which the measure was originally developed.[321] For example, measures developed to emphasize specific content areas would be expected to show greater change after treatment in those content areas.[314] Thus, the greater sensitivity to change in condition-specific PROs can likely be attributed to the strong content validity inherent in condition-specific measures. [30] As a result, using a combination of condition-specific and generic PRO measures may yield the most meaningful data.[30,32]

Minimally Important Differences and Changes

The difference between clinical versus statistical significance also merits consideration when selecting a PRO measure. Historically, research has relied upon tests of statistical significance to examine differences in PRO scores between patients or within patients over time. However, concerns arise regarding whether statistically significant differences truly reflect differences that would be perceived as important to the patient or the clinician. Consequently, attention has shifted to the concept of clinically significant differences in PRO scores. A variety of approaches to determining clinical significance have been proposed. For example, clinically significant change has been defined as "changes in patient functioning that are meaningful for individuals who undergo psychosocial or medical interventions."[322] Similarly, meaningful change is defined as ""one that results in a meaningful reduction in symptoms or improvement in function…" [from the patient perspective].[323] Minimally important differences (MIDs) represent a specific approach to clinical significance, and are defined as "…the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important."[324] Finally, minimum clinically important differences (MCIDs) comprise an even more specific category of MID and are defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management."[325]

The examination of clinically significant differences carries a number of important implications.[324] First, investigating clinically significant (versus statistically significant) differences in scores aids in the interpretation of PROs. Second, the focus on clinically significant differences also emphasizes the importance of the patient perspective, which may not be adequately captured when strictly looking at statistically significant differences. Third, the ability to look at clinically significant differences in PRO scores informs the evaluation of the success of a clinical intervention. Finally, in the context of clinical research, clinically significant differences can assist with sample size estimation.

Currently, no methodological "gold standard" exists for estimating MIDs;[323,326] however, two primary methods are currently in-use: the anchor-based method and the distribution-based method. The anchor-based method of establishing MIDs assesses the relationship between scores on the PRO and some independent measure which is interpretable.[324] Several options exist for the type of anchor selected when using the anchor-based method. First, clinical anchors which are correlated with the PRO measure at the $r \geq 0.30$ level may serve as appropriate anchors.[295,327] Clinical trial experience can be used to inform the selection of these clinical anchors,[328] which also enables the use of multiple clinical anchors.[329] Transition ratings represent another potential source of anchors when establishing MIDs. Transition ratings are within-person global ratings of change made by a patient.[327,330] However, due to concerns about validity, it is recommended that researchers examine the correlation between pre-and post-test PRO scores and the transition rating.[331] Between-person differences made by patients can also be used as anchors when establishing MIDs for PRO measures.[314,317] Additional sources for anchors when establishing MIDs include HRQL-related functional measures used by clinicians [327,330] and objective standards (e.g., hospital admissions, time away from work.[331] Although the

anchor-based method offers promise for establishing MIDs in PRO measures, several limitations should be considered. First, the transition rating approach to anchor selection is subject to recall bias on the part of the patient. [323] Second, global ratings may only account for some variance in PRO scores.[323] Third, the anchor based method does not take into consideration measurement precision of instrument.[323]

The distribution-based method represents the second method of establishing MIDs in PRO measures. The distribution-based method uses the statistical characteristics of the PRO scores when establishing MIDs.[324] Specifically, the distribution-based approach evaluates change in scores in relation to the probability that the change occurred at random.[323] As in the case of the anchor-based method, there are several methods available when applying a distribution-based approach to MID establishment. First, the t-test statistic has been used to establish MID when examining change over time.[323] However, given that this relies solely on statistical significance, it may not reflect change that is clinically meaningful and it is also subject to variation due to sample size.[323] Distribution-based methods may also be grounded in measurement precision and the standard error of mean (SEM). [323] Specifically, it has been suggested that the 1 SEM criterion can be used as an alternative to MID when assessing the magnitude of PRO score changes.[332] Sample variation, such as effect size and standardized response mean, constitutes another method for establishing MIDs using the distribution-based method.[323] When using this method, it is recommended that the effect size by specific to the population being studied.[330] Evidence suggests that MID estimates using sample variation are approximately half of a standard deviation.[333] Finally, reliable change constitutes another method of using the distribution-based approach to establishing MIDs.[323] Reliable change is based on the standard error of measurement difference (SEMD) and indicates how much the observed change exceeds fluctuations in an imprecise measure that are random in nature.[323] While the distribution-based approach serves as a possible alternative to the anchor-based methods, there is little consensus on the benchmarks for establishing changes that are clinically significant.[323]

Given limitations of the anchor- and distribution-based approaches, it is recommended that multiple methods and triangulation should be used to determine the MID.[295,323,333] Moreover, the final selection of MID values should be based on systematic review and an evaluation process such as the Delphi method.[295] MID values should also be informed by a stakeholder consensus, which includes patient engagement and input, about the extent of change considered to be meaningful. For example, there may be cases in which ability of scores over time is the desired outcome, such as in the case of interventions designed to preserve and prevent declines in functioning. Consequently, the application of the PRO is important in informing MID values, particularly when considering the contrasts between interventions for acute clinical conditions and interventions or support for long-term or chronic conditions. When considering MIDs for PRO measures, a single MID should not be applied to situation involving that particular PRO, given that MID varies by population/context.[295] Consequently, it is recommended that the distribution around the MID be provided rather than just a single MID value.[329] Finally, because the criteria for assessing clinically important change in individuals do not directly translate to evaluating clinically important group differences,[327] a useful strategy is to calculate the proportion of patients who experience a clinically significant change.[271,327]

Essential Conditions to Integrate PROs into the Electronic Health Record

Health information technology (HIT) has the potential to enable dramatic transformation in health care delivery, but the empirical research evidence base supporting its benefits is limited.[334]

E-health refers to health-related Internet applications that deliver a range of content, connectivity and clinical care.[8] This includes health information, online formularies, prescription refills, appointment scheduling, test results, advance care planning and health care proxy designation, and physician-patient communication.[335] Patient-Centered E-Health (PCEH) is an emerging discipline that is defined as the combination of three themes:[336]

- Patient-focus: PCEH applications are developed primarily based on needs and perspectives of patients.
- Patient-activity: PCEH application designs assume that patients can participate meaningfully in providing and consuming information about, and of interest to, them,
- Patient-empowerment: PCEH applications assume that patients want to, and are able to, control far-ranging aspects of their health care via a PCEH application.

Although e-health applications have become common, they tend to focus on the needs of health care providers and organizations. Patients desire a range of services to be brought online by their own health care provider.[337] However, there is little evidence about whether the services offered by providers are services that patients desire.[9] It is important that providers attend to patient acceptability factors.[9,338]

Measurement of PROs will constitute an important aspect of future stages of "meaningful use" of electronic health records (EHRs).[339,340] There is the potential for enhanced access by allowing entry directly from commonly used devices such as smart phones. Enabling clinical decision support by providing structured data directly into EHRs will permit PROs to (a) be used for tracking patient progress over time, or (b) use individual question responses to drive change in care plans or care processes concurrently thus improving outcomes over time. The use of a standardized instrument registered in an established code system (e.g., LOINC) enables EHRs to incorporate the instrument as an observation with a known set of responses using standard terminology (SNOMED-CT) or numerical responses. Each question in the standardized instrument can also be coded (structured) to drive changes based on those responses.

Unfortunately, in an updated systematic review of health information technology studies published during 2004-2007, PROs were not mentioned at all.[335]

The passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act creates a mix of incentives and penalties that will induce a large proportion of physicians and hospitals to move toward EHR systems by the end of this decade.[341] The discussion should now focus on whether HIT will support the models of care delivery that will help achieve broader policy goals: safer, more effective, and more efficient care.

Three features of EHRs are critical to enable accountable care organizations to succeed: interoperability and widespread health information exchange; automated, real-time quality and cost measurement; and smarter analytic capacities. Having a complete picture of the patient's care is a critical start, yet most EHRs are not interoperable and have limited data-sharing capabilities.[342] In summary, important issues include: a) the patient perspective (patients want to be involved "as a participant and partner in the flow of information" relating to their own health care[343]); b) clinical buy-in; c) compatibility with clinical flow; and d) meaningful use.

Examples. Health care centers are beginning to implement ways to use patient-reported information ("the voice of the patient") to provide higher quality care.[344] Three recent case studies (two in the U.S. and one in Sweden) are particularly informative to illustrate "lessons learned" about such initiatives.[344] The Dartmouth Spine Center collects health survey data from

patients before each visit, either at home or in the clinic. The data are summarized in a report and are available for use by the patients and clinicians to develop or modify the care plan, and to monitor results over time to guide treatment decisions. Longitudinal changes are incorporated into the report with each new assessment. The Karolinska University hospital (Stockholm, Sweden) developed a Swedish Rheumatology Quality registry in 1995 to improve the quality and value of care for people suffering from arthritis and other rheumatic diseases. Paper forms have now been replaced with a web-based system that makes use of real-time data provided by patients, clinicians and diagnostic tests. Longitudinal summaries of PRO measures and other health information are incorporated into graphical reports that are available to patients and providers. An electronic Health Risk Assessment has been integrated with an electronic health record at Group Health Cooperative in the State of Washington. Patients can complete PRO measures, make appointments, fill prescriptions, review health benefits, communicate with their providers, and get vetted health information. Customized reports are available to patients and providers.

Both patients and clinicians have generally favorable reactions to the patient-reported measurement systems implemented in these three very different health care settings. The information gathered helps to support patient-centered care by focusing attention on the health issues and outcomes that are important to patients. Although both patients and clinicians acknowledge that using PROs takes extra time for data collection, both groups report that it makes the care more effective and efficient. Key design principles to successful use of patient-reported measurement systems include fitting PRO measures into the flow of care, designing the systems with stakeholder engagement, merging PRO data with other types of data (clinician reports, medical records, claims), and engaging in continuous improvement of the systems based on users' experiences and new technology.

Other examples can be found in the use of PROs in the management of advanced cancer where the primary goals of care are to maximize symptom management and minimize treatment toxicity. Clinicians and patients often base treatment decisions on informal assessments of health-related quality of life (HRQL). Integrating formal HRQL assessment into treatment decision-making has the potential to improve patient-centered care for advanced cancer patients. Computer-based PRO assessment can reduce patient and administrative burden while enabling real-time scoring and presentation of HRQL data. Two pilot studies conducted with advanced lung cancer patients reported that the computer technology was acceptable and feasible for patients and physicians.[159,345] Patients felt that the HRQL questionnaire helped them focus on issues to discuss with their physicians, and physicians indicated that the HRQL report helped them to evaluate patient responses over time.

A new initiative in the Robert H. Lurie Comprehensive Cancer Center involves the development and implementation of patient-reported symptom assessment in Gynecologic Oncology clinics. Prior to clinic visits, outpatients complete instruments measuring fatigue, pain, physical function, depression and anxiety through the electronic health record (EHR) patient communication portal at home or in-clinic using an iPad. Results immediately populate the EHR. Severe symptoms trigger EHR notifications to providers. The EHR provides automated triage for psychosocial and nutritional care when indicated.

Selection of PROs that Meet the Recommended Characteristics for use in Performance Measures

A number of characteristics have been recommended when evaluating the appropriateness of a PRO for use in performance measures, as indicated in Table 4. Given that PROs are not

yet in widespread use in clinical practice, little is known about how best to aggregate these patient-level outcomes for the purpose of measuring performance of the health care entity. In spite of this, in order to accommodate the needs of patients with diverse linguistic, cultural, educational and functional skills, evidence is needed regarding the equivalence of multiple methods and modes of questionnaire administration. Additionally, scoring, analysis and reporting of PRO response data needs to be user-friendly and understandable to clinicians for use in real-time in clinical settings. Moreover, the timing of measurement must include pre-intervention in order to allow for measurement of responsiveness to change, to allow for risk adjustment, and to facilitate candidate screening for clinical intervention. In order to illustrate the application of these recommended characteristics when evaluating the appropriateness of a PRO for use as a performance measure, we provide the following example related to the evaluation of a PRO for use as a performance measure when evaluating the success of total hip arthroplasty.

Example of Applying Recommended Characteristics to Evaluate a Hip Osteoarthritis PRO for use in Performance Measurement. Total hip arthroplasty has emerged as an acceptable surgical treatment for individuals experiencing intractable pain and remarkable functional impairments for whom conservative treatment has yielded minimal improvement.[346,347 348,349] The most common indication for total hip arthroplasty is joint deterioration secondary to osteoarthritis.[350] Consequently, the aging of the population is likely to result in an increased demand for both primary, as well as revision total hip arthroplasty procedures.[351-353] Patient-reported outcomes have increasingly been included alongside more traditional indices of surgical outcome such as morbidity and mortality when evaluating the success of total hip arthroplasty as an intervention. With the increasing focus on patient-reported outcomes, such as functioning and quality of life, a widespread array of PROs have been developed and applied to the measurement of total hip arthroplasty outcomes.[350] Consequently, total hip arthroplasty provides a relevant context in which to review the use of recommended characteristics in the selection of PRO measures. Table 3 illustrates the application of important characteristics and best practices to evaluate and select a PRO for use as a performance measure for hip replacement outcomes. In this example, we illustrate the process of examining the characteristics of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), a PRO measure developed to examine pain, stiffness, and physical function in individuals with osteoarthritis.[354]

Table 4[1]. Important characteristics and best practices to evaluate and select PROs for use in performance measures[279,284]

	Characteristic	Specific issues to address for performance measures	Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[354] for use in hip arthroplasty
1.	**Conceptual and Measurement Model**		
	A PRO measure should have documentation defining and describing the concept(s) included and the intended population(s) for use.	• Target PRO concept should be a high priority for the health care system and patients. Patient engagement should define what is an important concept to patients. • Target PRO concept must be actionable in response to the healthcare intervention.	• Factorial validity of the physical function and pain subscales has been inadequate.[355]
	There should be documentation of how the concept(s) are organized into a measurement model, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts.		
2.	**Reliability**		
	The degree to which an instrument is free from random error.		
2a.	***Internal consistency*** *(multi-item scales)*	Classical Test Theory (CTT): ▪ reliability estimate ≥ 0.70 for group-level purposes ▪ reliability estimate ≥ 0.90 for individual-level purposes Item Response Theory: • item information curves that demonstrate precision [181] • a formula can be applied to estimate CTT reliability	• Cronbach alphas for the three subscales range from 0.86 to 0.98.[356-358]
2b.	***Reproducibility*** *(stability over time)* ▪ type of test-retest estimate depends on the response scale (dichotomous, nominal ordinal, interval, ratio)		• Test-retest reliability has been adequate for the pain and physical function subscales, but less adequate for the stiffness subscale.[358]
3.	**Validity**		

[1] This table is adapted from recommendations contained within a report from the Scientific Advisory Committee of the Medical Outcomes Trust and a report submitted to the PCORI Methodology Committee. The recommendations from these sources have been adapted to enhance relevance to PRO selection for performance measurement.

	Characteristic	Specific issues to address for performance measures	Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[354] for use in hip arthroplasty
	The degree to which the instrument reflects what it is supposed to measure.	• There are a limited number of PRO instruments that have been validated for performance measurement. • PRO instruments should include questions that are patient-centered.	
3a.	***Content Validity***		
	The extent to which a measure samples a representative range of the content.		
	A PRO measure should have evidence supporting its content validity, including evidence that patients and/or experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application.		• Development involved expert clinician input, and survey input from patients,[359] as well as a review of existing measures.
	Documentation of qualitative and/or quantitative methods used to solicit and confirm attributes (i.e., concepts measured by the items) of the PRO relevant to the measurement application.		
	Documentation of the characteristics of participants included in the evaluation (e.g., race/ethnicity, culture, age, socio-economic status, literacy).		
	Documentation of sources from which items were derived, modified, and prioritized during the PRO measure development process.		
	Justification for the recall period for the measurement application.		
3b.	***Construct and Criterion-related Validity***		
	A PRO measure should have evidence supporting its construct validity, including: • documentation of empirical findings that support predefined hypotheses on the expected associations among measures similar or dissimilar to the measured PRO • documentation of empirical findings that support predefined hypotheses of the expected differences in scores between "known" groups		• Patient ratings of satisfaction with arthroplasty were correlated with WOMAC scores in the expected direction.[22,360,361]
	A PRO measure should have evidence that shows the extent to which scores of the instrument are related to a criterion measure.		
3c.	***Responsiveness***		
	A PRO measure for use in longitudinal initiatives should have evidence of responsiveness, including empirical evidence of changes in scores consistent with predefined hypotheses regarding changes in the target population.	• If a PRO measure has cross-sectional data that provides sufficient evidence in regard to the reliability (internal consistency), content validity, and construct validity but has no data yet on responsiveness over time (i.e., ability of a	• Demonstrates adequate responsiveness and ability to detect change in response to clinical intervention.[362]

	Characteristic	Specific issues to address for performance measures	Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[354] for use in hip arthroplasty
		PRO measure to detect changes in the construct being measured over time), would you accept use of the PRO measure to provide valid data over time in a longitudinal study if no other PRO measure was available?	
		• Important to emphasize responsiveness because there is an expectation of consequences. Need to be able to demonstrate responsiveness if action is to be taken.	
		• PRO must be sensitive to detect change in response to the specific healthcare intervention	
4.	**Interpretability of Scores**		
	A PRO measure should have documentation to support interpretation of scores, including: • what low and high scores represent for the measured concept • representative mean(s) and standard deviation(s) in the reference population • guidance on the minimally important difference in scores between groups and/or over time that can be considered meaningful from the patient and/or clinical perspective	• If different PROs are used, it is important to establish a link or cross-walk between them. • Because the criteria for assessing clinically important change in individuals does not directly translate to evaluating clinically important group differences, [327] a useful strategy is to calculate the proportion of patients who experience a clinically significant change[271,327]	• Availability of population-based, age- and gender-normative values[363] • Availability of minimal clinically important improvement values[364] • Can be translated into a utility score for use in economic and accountability evaluations[365]
5.	**Burden**		
	The time, effort, and other demands on the respondent and the administrator.	• In a busy clinic setting, PRO assessment should be as brief as possible, and reporting should be done in real-time. • Patient engagement should inform what constitutes "burden."	• Short form available[366] • Average time to complete mobile phone WOMAC = 4.8 minutes[367]
6.	**Alternatives modes and methods of administration**	• The use of multiple modes and methods can be useful for diverse populations. However, there should be evidence regarding their	• Validated mobile phone and touchscreen based platforms[368,369]

	Characteristic	Specific issues to address for performance measures	Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[354] for use in hip arthroplasty
		equivalence.	
7.	**Cultural and language adaptations**	• The mode, method and question wording must yield equivalent estimates of PRO measures.	• Available in over 65 languages[370]
8.	**Electronic health records (EHR)**	Critical features: ▪ interoperability ▪ automated, real-time measurement and reporting ▪ sophisticated analytic capacities	▪ Electronic data capture may allow for integration within EHR[367]

V. Conclusion

Patient Reported Outcome (PRO) measures have reached a level of sophistication to enable their use in performance measures in the clinical setting. Attention to the many methodological considerations discussed in this paper will help produce meaningful, actionable results. Judicious use of a mixture of generic and condition-specific assessment, along with modern measurement methods such as item response theory, and the application of technology to enable standardized, equitable assessment across a range of patients, such as that applied in the development and validation of the PROMIS® instruments, can effectively shorten assessment time without compromising accuracy, meeting the demands of clinical application of PROs for performance measurement.

References

1. National Quality Forum (NQF). *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care.* Washington, DC: National Quality Forum; 2009.

2. US Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. 2009; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071975.pdf Accessed November 26, 2011.

3. Stephens RJ, Hopwood P, Girling DJ, Machin D. Randomized trials with quality of life endpoints: Are doctors' ratings of patients' physical symptoms interchangeable with patients' self- ratings? *Qual. Life Res.* 1997;6(3):225-236.

4. Justice AC, Rabeneck L, Hays RD, Wu AW, Bozzette SA, for the Outcomes Committee of the ACTG. Sensitivity, Specificity, Reliability, and Clinical Validity of Provider-Reported Symptoms: A Comparison With Self-Reported Symptoms. *J. Acquir. Immune Defic. Syndr.* 1999;21(2):126-133.

5. Basch E, Iasonos A, McDonough T, et al. Patient versus clinician symptom reporting using the National Cancer Institute Common Terminology Criteria for Adverse Events: results of a questionnaire-based study. *Lancet Oncol.* 2006;7(11):903-909.

6. Basch E, Jia X, Heller G, et al. Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes. *J. Natl. Cancer Inst.* 2009;101(23):1624-1632.

7. Basch E. The missing voice of patients in drug-safety reporting. *N Engl J Med.* 2010;362(10):865-869.

8. Maheu MM, Whitten P, Allen A. *E-Health, telehealth, and telemedicine: a guide to start-up and success.* San Francisco: Jossey-Bass; 2001.

9. Wilson EV, Lankton NK. Modeling patients' acceptance of provider-delivered e-health. *J. Am. Med. Inform. Assoc.* 2004;11(4):241-248.

10. Fayers P, Machin D. *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes.* 2nd ed. Chichester: John Wiley & Sons; 2007.

11. Bech P. Quality of life measurements in chronic disorders. *Psychother. Psychosom.* 1993;59(1):1-10.

12. Cella DF, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure. *J. Clin. Oncol.* 1993;11(3):570-579.

13. Guyatt GH. A taxonomy of health status instruments. *J. Rheumatol.* 1995;22(6):1188-1190.

14. Rothrock NE, Kaiser KA, Cella D. Developing a Valid Patient-Reported Outcome Measure. *Clin. Pharmacol. Ther.* 2011;90(5):737-742.

15. Osoba D. A taxonomy of the uses of health-related quality-of-life instruments in cancer care and the clinical meaningfulness of the results. *Med. Care.* 2002;40(6 Suppl):III31-38.

16. Benson T, Sizmur S, Whatling J, Arikan S, McDonald D, Ingram D. Evaluation of a new short generic measure of health status: howRu. *Inform. Prim. Care.* 2010;18(2):89-101.

17. Barry MJ, Fowler FJ, Jr., O'Leary MP, Bruskewitz RC, Holtgrewe HL, Mebust WK. Measuring disease-specific health status in men with benign prostatic hyperplasia. Measurement Committee of The American Urological Association. *Med. Care.* 1995;33(4 Suppl):AS145-155.

18. Ware JE, Jr., Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I.Conceptual Framework and Item Selection. *Med. Care.* 1992;30(6):473-483.

19. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: Development and final revision of a health status measure. *Med. Care.* 1981;19(8):787-805.

20. Caplan D, Hildebrandt N. *Disorders of Syntactic Comprehension.* Cambridge, Mass.: MIT Press; 1988.

21. Andresen EM, Rothenberg BM, Panzer R, Katz P, McDermott MP. Selecting a generic measure of health-related quality of life for use among older adults. A comparison of candidate instruments. *Eval. Health Prof.* 1998;21(2):244-264.

22. Bombardier C, Melfi CA, Paul J, et al. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. *Med. Care.* 1995;33(4 Suppl):AS131-144.

23. Lundgren-Nilsson A, Tennant A, Grimby G, Sunnerhagen KS. Cross-diagnostic validity in a generic instrument: an example from the Functional Independence Measure in Scandinavia. *Health Qual. Life Outcomes.* 2006;4:55.

24. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. *Med. Care.* 2007;45(5 Suppl 1):S3-S11.

25. Cella D, Riley W, Stone A, et al. Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005-2008. *J. Clin. Epidemiol.* 2011;63(11):1179-1194.

26. Cella D, Lai JS, Nowinski C, et al. Neuro-QOL: Brief Measures of Health-related Quality of Life for Clinical Research in Neurology. *Neurology.* 2012;Epub ahead of print.

27. Tulsky DS, Kisala PA, Victorson D, et al. Developing a Contemporary Patient-Reported Outcomes Measure for Spinal Cord Injury. *Arch. Phys. Med. Rehabil.* 2011;92(10, Supplement):S44-S51.

28. Cella D. *Manual of the Functional Assessment of Chronic Illness Therapy (FACIT Scales). Version 4* Elmhurst, IL: FACIT.org; 1997.

29. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. *Can. Med. Assoc. J.* 1986;134(8):889-895.

30. Eurich DT, Johnson JA, Reid KJ, Spertus JA. Assessing responsiveness of generic and specific health related quality of life measures in heart failure. *Health Qual. Life Outcomes.* 2006;4:89.

31. Huang IC, Hwang CC, Wu MY, Lin W, Leite W, Wu AW. Diabetes-specific or generic measures for health-related quality of life? Evidence from psychometric validation of the D-39 and SF-36. *Value Health.* 2008;11(3):450-461.

32. Krahn M, Bremner KE, Tomlinson G, Ritvo P, Irvine J, Naglie G. Responsiveness of disease-specific and generic utility instruments in prostate cancer patients. *Qual. Life Res.* 2007;16(3):509-522.

33. Cohen ME, Marino RJ. The tools of disability outcomes research functional status measures. *Arch. Phys. Med. Rehabil.* 2000;81(12 Suppl 2):S21-29.

34. Bombardier C, Tugwell P. Methodological considerations in functional assessment. *J. Rheumatol.* 1987;14 Suppl 15:6-10.

35. Gabel CP, Michener LA, Burkett B, Neller A. The Upper Limb Functional Index: development and determination of reliability, validity, and responsiveness. *J. Hand Ther.* 2006;19(3):328-348; quiz 349.

36. Hobart J, Kalkers N, Barkhof F, Uitdehaag B, Polman C, Thompson A. Outcome measures for multiple sclerosis clinical trials: relative measurement precision of the Expanded Disability Status Scale and Multiple Sclerosis Functional Composite. *Mult. Scler.* 2004;10(1):41-46.

37. Kaasa T, Loomis J, Gillis K, Bruera E, Hanson J. The Edmonton Functional Assessment Tool: preliminary development and evaluation for use in palliative care. *J. Pain Symptom Manage.* 1997;13(1):10-19.

38. Mausbach BT, Moore R, Bowie C, Cardenas V, Patterson TL. A review of instruments for measuring functional recovery in those diagnosed with psychosis. *Schizophr. Bull.* 2009;35(2):307-318.

39. Olarsch S. *Validity and responsiveness of the late-life function and disability instrument in a facility-dwelling population*, Boston University; 2008.

40. Litwin MS, Hays R, Fink A, Ganz PA, Leake B, Brook RH. The UCLA Prostate Cancer Index: Development, reliability, and validity of health-related quality of life measure. *Med. Care.* 1998;26(7):1002-1012.

41. Rosen R, Brown C, Heiman J, et al. The Female Sexual Function Index (FSFI): A multidimensional self-report instrument for the assessment of female sexual function. *J. Sex Marital Ther.* 2000;26(2):191-208.

42. Cleeland CS. Symptom burden: multiple symptoms and their impact as patient-reported outcomes. *J. Natl. Cancer Inst. Monogr.* 2007(37):16-21.

43. Smith E, Lai JS, Cella D. Building a measure of fatigue: the functional assessment of chronic illness therapy fatigue scale. *PM R.* 2010;2(5):359-363.

44. Yount SE, Choi SW, Victorson D, et al. Brief, Valid Measures of Dyspnea and Related Functional Limitations in Chronic Obstructive Pulmonary Disease (COPD). *Value Health.* 2011;14(2):307-315.

45. Amtmann D, Cook KF, Jensen MP, et al. Development of a PROMIS item bank to measure pain interference. *Pain.* 2010;150(1):173-182.

46. Centers for Disease Control and Prevention (CDC). Workplace Health Promotion-Glossary Terms. 2012; http://www.cdc.gov/workplacehealthpromotion/glossary/#H. Accessed September 25, 2012.

47. Oremus M, Hammill A, Raina P. *Health Risk Appraisal.* Rockville, MD: Agency for Healthcare Research and Quality; 2011.

48. Wellsource Inc. Scientific Validity. 2011; www.wellsource.com/scientific-validity.html. Accessed September 25, 2012.

49. Goetzel RZ, Ozminkowski RJ, Bruno JA, Rutter KR, Isaac F, Wang S. The long-term impact of Johnson & Johnson's Health & Wellness Program on employee health risks. *J. Occup. Environ. Med.* 2002;44(5):417-424.

50. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System. 2012; http://www.cdc.gov/brfss. Accessed September 11, 2012.

51. Centers for Disease Control and Prevention (CDC). *National Health and Nutrition Examination Survey, 2007-2008: Overview.* Hyattsville, MD: Centers for Disease Control National Center for Health Statistics; 2008.

52. Bonevski B, Campbell E, Sanson-Fisher RW. The validity and reliability of an interactive computer tobacco and alcohol use survey in general practice. *Addict. Behav.* 2010;35(5):492-498.

53. Couwenbergh C, van der Gaag RJ, Koeter M, de Ruiter C, van den Brink W. Screening for substance abuse among adolescents validity of the CAGE-AID in youth mental health care. *Subst. Use Misuse.* 2009;44(6):823-834.

54. Paxton AE, Strycker LA, Toobert DJ, Ammerman AS, Glasgow RE. Starting the conversation performance of a brief dietary assessment and intervention tool for health professionals. *Am. J. Prev. Med.* 2011;40(1):67-71.

55. Sallis R. Developing healthcare systems to support exercise: exercise as the fifth vital sign. *Br. J. Sports Med.* 2011;45(6):473-474.

56. Wong SL, Leatherdale ST, Manske SR. Reliability and validity of a school-based physical activity questionnaire. *Med. Sci. Sports Exerc.* 2006;38(9):1593-1600.

57. Morisky DE, Ang A, Krousel-Wood M, Ward HJ. Predictive validity of a medication adherence measure in an outpatient setting. *J. Clin. Hypertens.* 2008;10(5):348-354.

58. Agency for Healthcare Research and Quality. Notice Number: NOT-HS-05-005. Special Emphasis Notice: Research Priorities for the Agency for Healthcare Research and Quality. 2005 http://grants.nih.gov/grants/guide/notice-files/NOT-HS-05-005.html. Accessed June 25, 2012.

59. Institute of Medicine. *Crossing the quality chasm: a new health system for the 21st century.* Washington, D.C.: National Academy Press; 2001.

60. Hall JA, Dornan MC. Meta-analysis of satisfaction with medical care: Description of research domain and analysis of overall satisfaction levels. *Soc. Sci. Med.* 1988;27(6):637-644.

61. Lewis JR. Patient views on quality care in general practice: Literature review. *Soc. Sci. Med.* 1994;39(5):655-670.

62. Locker D, Dunt D. Theoretical and methodological issues in sociological studies of consumer satisfaction with medical care. *Soc. Sci. Med.* 1978;12:283-292.

63. Pascoe GC. Patient satisfaction in primary health care: A literature review and analysis. *Eval.Program.Plann.* 1983;6(3-4):185-210.

64. Williams B. Patient satisfaction: A valid concept? *Soc. Sci. Med.* 1994;38(4):509-516.

65. Shikiar R, Rentz AM. Satisfaction with medication: an overview of conceptual, methodologic, and regulatory issues. *Value Health.* 2004;7(2):204-215.

66. Linder-Pelz SU. Toward a theory of patient satisfaction. *Soc. Sci. Med.* 1982;16(5):577-582.

67. Oberst MT. Patients' perceptions of care. Measurement of quality and satisfaction. *Cancer.* 1984;53(10):2366-2375.

68. National Quality Forum (NQF). *Safe Practices for Better Healthcare–2010 Update.* Washington, D.C.: National Quality Forum; 2010.

69. Ware JE, Jr., Snyder MK, Wright WR, Davies AR. Defining and measuring patient satisfaction with medical care. *Eval. Program Plann.* 1983;6(3-4):247-263.

70. Cella D, Bonomi A, Leslie WT, VonRoenn J, Tchekmedyian NS. Quality of life and nutritional well-being: measurement and relationship. *Oncology.* 1993;7(11, Suppl):S105-S111.

71. Rubin HR, Gandek B, Rogers WH, Kosinski M, McHorney CA, Ware JE, Jr. Patients' Ratings of Outpatient Visits in Different Practice Settings. Results from the Medical Outcomes Study. *JAMA.* 1993;270(7):835-840.

72. Graham J. Foundation for accountability(FACCT): a major new voice in the quality debate. In: Boyle J, ed. *1997 Medical Outcomes & Guidelines Sourcebook : a progress*

report and resource guide on medical outcomes research and practice guidelines : developments, data, and documentation. New York: Faulkner & Gray; 1996.

73. Hays RD, Davies AR, Ware JE. Scoring the Medical Outcomes Study Patient Satisfaction Questionnaire: PSQ-III. Unpublished work 1987.

74. Moinpour CM. Assessment of quality of life in clinical trials. *Quality of life assesment in cancer clinical trials. Report of the Workshop on Quality of Life Research in Cancer Clinical Trials, July 16-17, 1990*. Bethesda, MD: U.S. Department of Health and Human Services; 1991.

75. Williams S. Consumer satisfaction surveys: health plan report cards to guide consumers in selecting benefit programs. In: Boyle J, ed. *1997 Medical Outcomes & Guidelines Sourcebook : a progress report and resource guide on medical outcomes research and practice guidelines : developments, data, and documentation*. New York: Faulkner & Gray; 1996.

76. Speight J. Assessing patient satisfaction: concepts, applications, and measurement. *Value Health.* 2005;8 Suppl 1:S6-8.

77. Epstein LH, Cluss PA. A behavioral medicine perspective on adherence to long-term medical regimens. *J. Consult. Clin. Psychol.* 1982;50(6):950-971.

78. Sherbourne CD, Hays RD, Ordway L, DiMatteo MR, Kravitz RL. Antecedents of adherence to medical recommendations: Results from the Medical Outcomes Study. *J. Behav. Med.* 1992;15(5):447-468.

79. Hays RD, Kravitz RL, Mazel RM, et al. The impact of patient adherence on health outcomes for patients with chronic disease in the Medical Outcomes Study. *J. Behav. Med.* 1994;17(4):347-360.

80. Hirsh AT, Atchison JW, Berger JJ, et al. Patient Satisfaction With Treatment for Chronic Pain: Predictors and Relationship to Compliance. *Clin. J. Pain.* 2005;21(4):302-310.

81. Ickovics JR, Meisler AW. Adherence in AIDS clinical trials: a framework for clinical research and clinical care. *J. Clin. Epidemiol.* 1997;50(4):385-391.

82. Kincey J, Bradshaw P, Ley P. Patients' satisfaction and reported acceptance of advice in general practice. *J. R. Coll. Gen. Pract.* 1975;25(157):558-566.

83. Augustin M, Reich C, Schaefer I, Zschocke I, Rustenbach SJ. Development and validation of a new instrument for the assessment of patient-defined benefit in the treatment of acne. *Journal der Deutschen Dermatologischen Gesellschaft.* 2008;6(2):113-120.

84. Blais MA. Development of an inpatient treatment alliance scale. *J. Nerv. Ment. Dis.* 2004;192(7):487-493.

85. Brod M, Christensen T, Bushnell D. Maximizing the value of validation findings to better understand treatment satisfaction issues for diabetes. *Qual. Life Res.* 2007;16(6):1053-1063.

86. Flood EM, Beusterien KM, Green H, et al. Psychometric evaluation of the Osteoporosis Patient Treatment Satisfaction Questionnaire (OPSAT-Q), a novel measure to assess satisfaction with bisphosphonate treatment in postmenopausal women. *Health Qual. Life Outcomes.* 2006;4:42.

87. Hudak PL, Hogg-Johnson S, Bombardier C, McKeever PD, Wright JG. Testing a new theory of patient satisfaction with treatment outcome. *Med. Care.* 2004;42(8):726-739.

88. Kumar RN, Kirking DM, Hass SL, et al. The association of consumer expectations, experiences and satisfaction with newly prescribed medications. *Qual. Life Res.* 2007;16(7):1127-1136.

89. Pouchot J, Trudeau E, Hellot SC, Meric G, Waeckel A, Goguel J. Development and psychometric validation of a new patient satisfaction instrument: the osteoARthritis Treatment Satisfaction (ARTS) questionnaire. *Qual. Life Res.* 2005;14(5):1387-1399.

90. Taback NA, Bradley C. Validation of the genital herpes treatment satisfaction questionnaire (GHerpTSQ) in status and change versions. *Qual. Life Res.* 2006;15(6):1043-1052.

91. Cella DF. Quality of life: The concept. *J. Palliat. Care.* 1992;8(3):8-13.

92. Wagner EH. Chronic disease management: what will it take to improve care for chronic illness? *Eff. Clin. Pract.* 1998;1(1):2-4.

93. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. *J. Gen. Intern. Med.* 2012;27(5):520-526.

94. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the Patient Activation Measure (PAM): Conceptualizing and Measuring Activation in Patients and Consumers. *Health Serv. Res.* 2004;39(4p1):1005-1026.

95. Hibbard JH. Using Systematic Measurement to Target Consumer Activation Strategies. *Med. Care Res. Rev.* 2009;66(1 suppl):9S-27S.

96. Hibbard JH, Mahoney ER, Stock R, Tusler M. Do Increases in Patient Activation Result in Improved Self-Management Behaviors? *Health Serv. Res.* 2007;42(4):1443-1463.

97. Lake T, Kvan C, Gold M. Literature Review: Using Quality Information for Health Care Decisions and Quality Improvement. *Mathematica Policy Research.* 2005;Reference No. 6110-230.

98. Schneider EC, Zaslavsky AM, Landon BE, Lied TR, Sheingold S, Cleary PD. National quality monitoring of Medicare health plans: the relationship between enrollees' reports and the quality of clinical care. *Med. Care.* 2001;39(12):1313-1325.

99. Browne K, Roseman D, Shaller D, Edgman-Levitan S. Analysis & commentary. Measuring patient experience as a strategy for improving primary care. *Health Aff. (Millwood).* 2010;29(5):921-925.

100. Cella DF, Lloyd SR. Data collection strategies for patient-reported information. *Qual. Manag. Health Care.* 1994;2(4):28-35.

101. Fowler FJ, Jr., Spilker B. Data Collection Methods. *Quality of Life and Pharmacoeconomics in Clinical Trials.* Vol 2nd. Philadelphia: Lippincott-Raven Publishers; 1996.

102. Naughton MJ, Shumaker SA, Anderson RT, Czajkowski SM, Spilker B. Psychological aspects of health-related quality of life measurement: tests and scales. *Quality of Life and Pharmacoeconomics in Clinical Trials.* Vol 2nd. Philadelphia: Lippincott-Raven; 1996.

103. Sneeuw KC, Sprangers MA, Aaronson NK. 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(11):1130-1143.

104. Eiser C, Morse R. A review of measures of quality of life for children with chronic illness. *Arch. Dis. Child.* 2001;84(3):205-211.

105. Eiser C, Morse R. Quality-of-life measures in chronic diseases of childhood. *Health Technol. Assess.* 2001;5(4):1-157.

106. Weinfurt KP, Trucco SM, Willke RJ, Schulman KA. Measuring agreement between patient and proxy responses to multidimensional health-related quality-of-life measures in clinical trials. An application of psychometric profile analysis. *J. Clin. Epidemiol.* 2002;55(6):608-618.

107. Andresen EM, Vahle VJ, Lollar D. Proxy reliability: Health-related quality of life (HRQoL) measures for people with disability. *Qual. Life Res.* 2001;10(7):609-619.

108. Hart T, Whyte J, Polansky M, et al. Concordance of patient and family report of neurobehavioral symptoms at 1 year after traumatic brain injury. *Arch. Phys. Med. Rehabil.* 2003;84(2):204-213.

109. Matziou V, Perdikaris P, Feloni D, Moshovi M, Tsoumakas K, Merkouris A. Cancer in childhood: Children's and parents' aspects for quality of life. *Eur J Oncol Nurs.* 2008;12(3):209-216.

110. Matziou V, Tsoumakas K, Perdikaris P, Feloni D, Moschovi M, Merkouris A. Corrigendum to: "Cancer in childhood: Children's and parents' aspects for quality of life" [Eur J Oncol Nurs 12 (2008) 209-216] (DOI:10.1016/j.ejon.2007.10.005). *Eur J Oncol Nurs.* 2009;13(5).

111. Oczkowski C, O'Donnell M. Reliability of Proxy Respondents for Patients With Stroke: A Systematic Review. *J. Stroke Cerebrovasc. Dis.* 2010;19(5):410-416.

112. Brown-Jacobsen AM, Wallace DP, Whiteside SPH. Multimethod, multi-informant agreement, and positive predictive value in the identification of child anxiety disorders using the SCAS and ADIS-C. *Assessment.* 2011;18(3):382-392.

113. Agnihotri K, Awasthi S, Singh U, Chandra H, Thakur S. A study of concordance between adolescent self-report and parent-proxy report of health-related quality of life in school-going adolescents. *J. Psychosom. Res.* 2010;69(6):525-532.

114. Dorman PJ, Waddell F, Slattery J, Dennis M, Sandercock P. Are proxy assessments of health status after stroke with the EuroQol questionnaire feasible, accurate, and unbiased? *Stroke.* 1997;28(10):1883-1887.

115. Duncan PW, Lai SM, Tyler D, Perera S, Reker DM, Studenski S. Evaluation of proxy responses to the Stroke Impact Scale. *Stroke.* 2002;33(11):2593-2599.

116. Ostbye T, Tyas S, McDowell I, Koval J. Reported activities of daily living: agreement between elderly subjects with and without dementia and their caregivers. *Age Ageing.* 1997;26(2):99-106.

117. Sneeuw KC, Aaronson NK, de Haan RJ, Loeb JM. Assessing quality of life after stroke. The value and limitations of proxy ratings. *Stroke.* 1997;28(8):1541-1549.

118. Morrow AM, Hayen A, Quine S, Scheinberg A, Craig JC. A comparison of doctors', parents' and children's reports of health states and health-related quality of life in children with chronic conditions. *Child Care Health Dev.* 2012;38(2):186-195.

119. White-Koning M, Arnaud C, Dickinson HO, et al. Determinants of child-parent agreement in quality-of-life reports: a European study of children with cerebral palsy. *Pediatrics.* 2007;120(4):804-814.

120. Upton P, Lawford J, Eiser C. Parent-child agreement across child health-related quality of life instruments: a review of the literature. *Qual. Life Res.* 2008;17(6):895-913.

121. Hilari K, Owen S, Farrelly SJ. Proxy and self-report agreement on the Stroke and Aphasia Quality of Life Scale-39. *J. Neurol. Neurosurg. Psychiatry.* 2007;78(10):1072-1075.

122. Lynn Snow A, Cook KF, Lin P-S, Morgan RO, Magaziner J. Proxies and Other External Raters: Methodological Considerations. *Health Serv. Res.* 2005;40(5p2):1676-1693.

123. Groves RM. *Survey methodology.* 2nd ed. Hoboken, NJ: J. Wiley; 2009.

124. Selltiz C, Wrightsman LS, Cook SW. *Research Methods in Social Relations.* New York: Holt, Rinehart and Winston; 1976.

125. Edwards AL. *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts; 1957.

126. Crowne DP, Marlowe D. *The approval motive: Studies in evaluative dependence.* New York: Wiley; 1964.

127. Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J. Public Health.* 2005;27(3):281-291.

128. Anderson JP, Bush JW, Berry CC. Classifying function for health outcome and quality-of-life evaluation. Self- versus interviewer modes. *Med. Care.* 1986;24(5):454-469.

129. Cook DJ, Guyatt GH, Juniper E, et al. Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma. *J. Clin. Epidemiol.* 1993;46(6):529-534.

130. McHorney CA, Kosinski M, Ware JE, Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey. *Med. Care.* 1994;32(6):551-567.

131. Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Med. Care.* 2004;42(3):281-289.

132. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? *J. Clin. Epidemiol.* 1996;49(2):135-140.

133. Chambers LW, Haight M, Norman G, MacDonald L. Sensitivity to change and the effect of mode of administration on health status measurement. *Med. Care.* 1987;25(6):470-480.

134. Wu AW, Jacobson DL, Berzon RA, et al. The effect of mode of administration on medical outcomes study health ratings and EuroQol scores in AIDS. *Qual. Life Res.* 1997;6(1):3-10.

135. Teresi JA. Overview of quantitative measurement methods: equivalence, invariance, and differential item functioning in health applications. *Med. Care.* 2006;44(11 Suppl 3):S39-S49.

136. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Med. Care.* 2006;44(11 Suppl 3):S152-S170.

137. Borsboom D. When does measurement invariance matter? *Med. Care.* 2006;44(11 Suppl 3):S176-S181.

138. Hambleton RK. Good practices for identifying differential item functioning. *Med. Care.* 2006;44(11 Suppl 3):S182-S188.

139. McHorney CA, Fleishman JA. Assessing and Understanding Measurement Equivalence in Health Outcome Measures: Issues for Further Quantitative and Qualitative Inquiry. *Med. Care.* 2006;44(11 Suppl 3):S205-S210.

140. Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value Health.* 2009;12(4):419-429.

141. Hahn E, Cella D, Dobrez D, et al. The Talking Touchscreen: a new approach to outcomes assessment in low literacy. *Psychooncology.* 2004;13(2):86-95.

142. Hahn EA, Cella D, Dobrez DG, et al. Quality of life assessment for low literacy Latinos: A new multimedia program for self-administration. *J. Oncol. Manag.* 2003;12(5):9-12.

143. Greist JH, Van Cura LJ, Erdman HP. Computer interview questionnaires for drug use/abuse. In: Lettieri DJ, National Institute on Drug Abuse, eds. *Predicting adolescent drug abuse : a review of issues, methods and correlates*. Rockville, Md.; Washington: U.S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute on Drug Abuse; 1975:164-174.

144. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value Health.* 2008;11(2):322-333.

145. Dalal AA, Nelson L, Gilligan T, McLeod L, Lewis S, DeMuro-Mercon C. Evaluating patient-reported outcome measurement comparability between paper and alternate versions, using the lung function questionnaire as an example. *Value Health.* 2011;14(5):712-720.

146. Abernethy AP, Herndon JE, Wheeler JL, et al. Improving health care efficiency and quality using tablet personal computers to collect research-quality, patient-reported data. *Health Serv. Res.* 2008;43(6):1975-1991.

147. Lohr K, Zebrack B. Using patient-reported outcomes in clinical practice: challenges and opportunities. *Qual. Life Res.* 2009;18(1):99-107.

148. Abernethy AP, Zafar SY, Wheeler JL, Lyerly HK, Ahmad A, Reese JB. Electronic patient-reported data capture as a foundation of rapid learning cancer care. *Med. Care.* 2010;48(6 SUPPL.):S32-S38.

149. Dudgeon D, King S, Howell D, et al. Cancer Care Ontario's experience with implementation of routine physical and psychological symptom distress screening. *Psychooncology.* 2012;21(4):357-364.

150. Gilbert JE, Howell D, King S, et al. Quality Improvement in Cancer Symptom Assessment and Control: The Provincial Palliative Care Integration Project (PPCIP). *J. Pain Symptom Manage.* 2012;43(4):663-678.

151. Snyder CF, Jensen R, Courtin SO, Wu AW. PatientViewpoint: a website for patient-reported outcomes assessment. *Qual. Life Res.* 2009;18(7):793-800.

152. Velikova G, Booth L, Smith AB, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial. *J. Clin. Oncol.* 2004;22(4):714-724.

153. Detmar SB, Muller MJ, Schornagel JH, Wever LD, Aaronson NK. Health-related quality-of-life assessments and patient-physician communication: A randomized controlled trial. *JAMA.* 2002;288(23):3027-3034.

154. Velikova G, Brown JM, Smith AB, Selby PJ. Computer-based quality of life questionnaires may contribute to doctor-patient interactions in oncology. *Br. J. Cancer.* 2002;86(1):51-59.

155. Suh SY, Leblanc TW, Shelby RA, Samsa GP, Abernethy AP. Longitudinal patient-reported performance status assessment in the cancer clinic is feasible and prognostic. *J Oncol Pract.* 2011;7(6):374-381.

156. Fihn SD, Bucher JB, McDonell M. Collaborative care intervention for stable ischemic heart disease. *Arch. Intern. Med.* 2011;171(16):1471-1479.

157. Fihn SD, McDonell MB, Diehr P, et al. Effects of sustained audit/feedback on self-reported health status of primary care patients. *Am. J. Med.* 2004;116(4):241-248.

158. Au DH, McDonell MB, Martin DC, Fihn SD. Regional Variations in Health Status. *Med. Care.* 2001;39(8):879-888.

159. Chang CH, Cella D, Masters GA, et al. Real-time clinical application of quality-of-life assessment in advanced lung cancer. *Clin Lung Cancer.* 2002;4(2):104-109.

160. Wright EP, Selby PJ, Crawford M, et al. Feasibility and compliance of automated measurement of quality of life in oncology practice. *J. Clin. Oncol.* 2003;21(2):374-382.

161. Valderas J, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual. Life Res.* 2008;17(2):179-193.

162. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *J. Eval. Clin. Pract.* 2006;12(5):559-568.

163. Mullen KH, Berry DL, Zierler BK. Computerized symptom and quality-of-life assessment for patients with cancer part II: acceptability and usability. *Oncol. Nurs. Forum.* 2004;31(5):E84-E89.

164. Jones JB, Snyder CF, Wu AW. Issues in the design of Internet-based systems for collecting patient-reported outcomes. *Qual. Life Res.* 2007;16(8):1407-1417.

165. Cleeland CS, Wang XS, Shi Q, et al. Automated symptom alerts reduce postoperative symptom severity after cancer surgery: a randomized controlled clinical trial. *J. Clin. Oncol.* 2011;29(8):994-1000.

166. Basch E, Artz D, Iasonos A, et al. Evaluation of an online platform for cancer patient self-reporting of chemotherapy toxicities. *J. Am. Med. Inform. Assoc.* 2007;14(3):264-268.

167. Hardwick ME, Pulido PA, Adelson WS. The use of handheld technology in nursing research and practice. *Orthop. Nurs.* 2007;26(4):251-255.

168. Bollen K, Lennox R. Conventional wisdom on measurement: A structural equation perspective. *Psychol. Bull.* 1991;110(2):305-314.

169. MacCallum RC, Browne MW. The use of causal indicators in covariance structure models: some practical issues. *Psychol. Bull.* 1993;114(3):533-541.

170. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. *Qual. Life Res.* 1997;6(2):139-150.

171. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. *Qual. Life Res.* 1997;6(5):393-406.

172. Sebille V, Hardouin J-B, Le Neel T, et al. Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients--a simulation study. *BMC Med. Res. Methodol.* 2010;10:24.

173. Bjorner JB, Chang C-H, Thissen D, Reeve BB. Developing tailored instruments: Item banking and computerized adaptive assessment. *Qual. Life Res.* 2007;16(Suppl1):95-108.

174. Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health Serv. Res.* 2005;40(5 Pt 2):1694-1711.

175. Cook KF, Teal CR, Bjorner JB, et al. IRT health outcomes data analysis project: an overview and summary. *Qual. Life Res.* 2007;16 Suppl 1:121-132.

176. Coster W, Ludlow L, Mancini M. Using IRT variable maps to enrich understanding of rehabilitation data. *J. Outcome Meas.* 1999;3(2):123-133.

177. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual. Life Res.* 2007;16 Suppl 1:5-18.

178. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Qual. Life Res.* 2007;16 Suppl 1:187-194.

179. Fries JF, Bruce B, Cella D. The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clin. Exp. Rheumatol.* 2005;23(S38):S33-S37.

180. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br. J. Clin. Psychol.* 2007;46(Pt 1):1-18.

181. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med. Care.* 2007;45(5 Suppl 1):S22-S31.

182. Nunnally JC, Bernstein IH. *Psychometric Theory.* New York: McGraw-Hill, Inc.; 1994.

183. Fleiss JL. *The Design and Analysis of Clinical Experiments.* New York: John Wiley & Sons; 1986.

184. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley 1968.

185. Allen MJ, Yen WM. *Introduction to measurement theory.* Monterey, CA: Brooks/Cole Publishing; 1979.

186. DeVellis RF. Classical test theory. *Med. Care.* 2006;44(11 Suppl 3):S50-S59.

187. DeVellis RF. *Scale development theory and applications.* Thousand Oaks, CA: Sage; 2003.

188. Martinez-Martin P. Composite rating scales. *J. Neurol. Sci.* 2010;289(1-2):7-11.

189. Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use.* New York: Oxford University Press; 2003.

190. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory.* Newbury Park, CA: SAGE Publications, Inc.; 1991.

191. Hambleton RK. Emergence of item response modeling in instrument development and data analysis. *Med. Care.* 2000;38(9 Suppl):II60-II65.

192. van der Linden WJ, Hambleton RK. *Handbook of Modern Item Response Theory.* New York Springer-Verlag; 1997.

193. Wright BD, Masters GN. *Rating scale analysis: Rasch measurement.* Chicago: MESA Press; 1985.

194. Cook KF, Monahan PO, McHorney CA. Delicate balance between theory and practice: Health status assessment and Item Response Theory. *Med. Care.* 2003;41(5):571-574.

195. McHorney CA, Cohen AS. Equating health status measures with Item Response Theory: Illustrations with functional status items. *Med. Care.* 2000;38(9 Suppl):1143-1159.

196. Dorans N. Comparing or combining scores from multiple instruments: instrument linking (equating). Paper presented at: Advances in Health Outcomes Measurement2004; Bethesda, MD.

197. *Quality First: Better Health Care for All Americans. Final Report of the President's Advisory Commission on Consumer Protection and Quality in the Health Care Industry.* Washington, DC: US Government Printing Office;1998.

198. Hahn EA, Cella D. Health outcomes assessment in vulnerable populations: Measurement challenges and recommendations. *Arch. Phys. Med. Rehabil.* 2003;84(Suppl 2):S35-S42.

199. United States Agency for Healthcare Research Quality. *National healthcare disparities report 2010* Rockville, Md.: Agency for Healthcare Research and Quality 2011.

200. Hahn E, Cella D, Dobrez D, et al. The impact of literacy on health-related quality of life measurement and outcomes in cancer outpatients. *Qual. Life Res.* 2007;16(3):495-507.

201. Kirsch I, Jungeblut A, Jenkins L, Kolstad A. *Adult literacy in America: A first look at the results of the National Adult Literacy Survey.* Washington, DC: National Center for Education Statistics, U.S. Department of Education; 1993.

202. Kutner M, National Center for Education Statistics. *Literacy in everyday life: results from the 2003 National Assessment of Adult Literacy (NCES 2007-480). U.S. Department of Education.* Washington, DC: National Center for Education Statistics; 2007.

203. U.S. Department of Health and Human Services. *Healthy People 2010: Understanding and Improving Health.* Washington, D.C.: U.S. Government Printing Office;2000. 2nd ed.

204. Committee on Health Literacy, Nielsen-Bohlman L, Panzer AM, Kindig DA. *Health Literacy: A Prescription to End Confusion.* Washington, D.C.: The National Academies Press; 2004.

205. Berkman ND, Sheridan SL, Donahue KE, et al. *Health Literacy Interventions and Outcomes: An Updated Systematic Review, Executive Summary, Evidence Report/Technology Assessment No. 199.* Rockville, MD: Agency for Healthcare Research and Quality; March 2011. Publication Number 11-E006-1

206. Kutner M, Greenberg E, Jin Y, Paulsen C. *The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy (NCES 2006–483).* Washington, DC: National Center for Education Statistics: U.S. Department of Education; 2006.

207. Baker DW, Gazmararian JA, Williams MV, et al. Functional health literacy and the risk of hospital admission among Medicare managed care enrollees. *Am. J. Public Health.* 2002;92(8):1278-1283.

208. DeWalt DA, Berkman ND, Sheridan S, Lohr KN, Pignone MP. Literacy and health outcomes: A systematic review of the literature. *J. Gen. Intern. Med.* 2004;19(12):1228-1239.

209. Rudd RE, Anderson JE, Oppenheimer S, Nath C. Health literacy: an update of public health and medical literature. In: Comings JP, Garner B, Smith CA, eds. *Review of adult learning and literacy.* Vol 7. Mahwah: Lawrence Erlbaum Associates; 2007:175-204.

210. Macabasco-O'Connell A, DeWalt D, Broucksou K, et al. Relationship Between Literacy, Knowledge, Self-Care Behaviors, and Heart Failure-Related Quality of Life Among Patients With Heart Failure. *J. Gen. Intern. Med.* 2011:1-8.

211. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association. Health literacy: Report of the Council on Scientific Affairs. *JAMA.* 1999;281(6):552-557.

212. Parikh NS, Parker RM, Nurss JR, Baker DW, Williams MV. Shame and health literacy: The unspoken connection. *Patient Educ. Couns.* 1996;27(1):33-39.

213. Baker DW, Parker RM, Williams MV, Coates WC, Pitkin K. Use and effectiveness of interpreters in an emergency department. *JAMA.* 1996;274(10):783-788.

214. Lennon C, Burdick H. The Lexile Framework As An Approach For Reading Measurement And Success. 2004; http://www.lexile.com/m/uploads/whitepapers/Lexile-Reading-Measurement-and-Success-0504_MetaMetricsWhitepaper.pdf. Accessed January 25, 2011.

215. Klare GR. *The measurement of readability.* Ames: Iowa State University Press; 1963.

216. Liberman IY, Mann VA, Shankweiler D, Werfelman M. Children's memory for recurring linguistic and nonlinguistic material in relation to reading ability. *Cortex.* 1982;18(3):367-375.

217. Shankweiler D, Crain S. Language mechanisms and reading disorder: a modular approach. *Cognition.* 1986;24(1-2):139-168.

218. Crain S, Shankweiler D. Syntactic Complexity and Reading Acquisition. In: Davidson A, Green GM, eds. *Linguistic complexity and text comprehension: readability issues reconsidered.* Vol Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1988:167-192.

219. Brach C, Keller D, Hernandez LM, et al. *Ten Attributes of Health Literate Health Care Organizations.* Washington, D.C.: National Academies Press; 2012.

220. U.S. Department of Health and Human Services OoMH, ,. *National Standards for Culturally and Linguistically Appropriate Services in Health Care.* Washington, DC: US Department of Health and Human Services; 2001.

221. Drasgow F, Kanfer R. Equivalence of psychological measurement in heterogeneous populations. *J. Appl. Psychol.* 1985;70:662-680.

222. Hui CH, Triandis HC. Measurement in cross-cultural psychology. *J Cross Cult Psychol.* 1985;16(2):131-152.

223. Angel R, Thoits P. The impact of culture on the cognitive structure of illness. *Cult. Med. Psychiatry.* 1987;11 23-52.

224. Bullinger M, Anderson R, Cella D, Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Qual. Life Res.* 1993;2(6):451-459.

225. Hayes RP, Baker DW. Methodological problems in comparing English-speaking and Spanish-speaking patients' satisfaction with interpersonal aspects of care. *Med. Care.* 1998;36(2):230-236.

226. Bjorner JB, Thunedborg K, Kristensen TS, Modvig J, Bech P. The Danish SF-36 Health Survey: Translation and preliminary validity studies. *J. Clin. Epidemiol.* 1998;51(11):991-999.

227. Hunt SM. Cross-Cultural Comparability of Quality of Life Measures. *Drug Inf. J.* 1993;27(2):395-400.

228. Atkinson MJ, Lennox RD. Extending basic principles of measurement models to the design and validation of Patient Reported Outcomes. *Health Qual. Life Outcomes.* 2006;4(1):65.

229. da Mota Falcao D, Ciconelli RM, Ferraz MB. Translation and cultural adaptation of quality of life questionnaires: an evaluation of methodology. *J. Rheumatol.* 2003;30(2):379-385.

230. Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: The universalist approach. *Qual. Life Res.* 1998;7(4):323-335.

231. Herdman M, Fox-Rushby J, Badia X. Equivalence and the translation and adaptation of health-related quality of life questionnaires. *Qual. Life Res.* 1997;6(3):237-247.

232. Wild D, Eremenco S, Mear I, et al. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. *Value Health.* 2009;12(4):430-440.

233. Wild D, Grove A, Martin M, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient reported outcomes(PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health.* 2005;8(2):94-104.

234. Acquadro C, Conway K, Hareendran A, Aaronson N. Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value Health.* 2008;11(3):509-521.

235. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* 2000;25(24):3186-3191.

236. Dewolf L, Koller M, Velikova G, Johnson C, Scott N, Bottomley A. EORTC Quality of Life Group: Translation Procedure. 2009; 3rd:http://groups.eortc.be/qol/downloads/translation_manual_2009.pdf. Accessed November 26, 2011.

237. Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Eval. Health Prof.* 2005;28(2):212-232.

238. Sperber AD. Translation and validation of study instruments for cross-cultural research. *Gastroenterology.* 2004;126(1 Suppl 1):S124-128.

239. Ware JE, Jr., Keller SD, Gandek B, Brazier JE, Sullivan M. Evaluating translations of health status questionnaires. Methods from the IQOLA project. International Quality of Life Assessment. *Int. J. Technol. Assess. Health Care.* 1995;11(3):525-551.

240. Centers for Disease Control and Prevention. Prevalence and most common causes of disability among adults - United States, 2005. *MMWR Morb. Mortal. Wkly. Rep.* 2009;58(16):421-426.

241. National Council on Disability. The current state of health care for people with disabilities. 2009; http://purl.fdlp.gov/GPO/gpo3755.

242. Agency for Healthcare Research and Quality. Developing Quality of Care Measures for People with Disabilities: Summary of Expert Meeting. AHRQ Publication No. 10-0103. 2010; http://www.ahrq.gov/populations/devqmdis/.

243. North Carolina State University College of Design. Center for Universal Design. 2008; http://www.ncsu.edu/project/design-projects/udi/.

244. Story MF. Maximizing Usability: The Principles of Universal Design. *Assist. Technol.* 1998;10(1):4-12.

245. Section 508 of the Rehabilitation Act, as amended by the Workforce Investment Act of 1998 (P.L. 105-220). 1998; http://www.section508.gov/. Accessed February 20, 2010.

246. Harniss M, Amtmann D, Cook D, Johnson K. Considerations for Developing Interfaces for Collecting Patient-Reported Outcomes That Allow the Inclusion of Individuals With Disabilities. *Med. Care.* 2007;45(5 Suppl 1):S48-S54.

247. Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc. Sci. Med.* 1999;48(11):1531-1548.

248. Nolte S, Elsworth GR, Sinclair AJ, Osborne RH. Tests of measurement invariance failed to support the application of the "then-test". *J. Clin. Epidemiol.* 2009;62(11):1173-1180.

249. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual. Life Res.* 2002;11(3):207-221.

250. Brossart DF, Clay DL, Willson VL. Methodological and statistical considerations for threats to internal validity in pediatric outcome data: response shift in self-report outcomes. *J. Pediatr. Psychol.* 2002;27(1):97-107.

251. Schwartz CE. Applications of response shift theory and methods to participation measurement: a brief history of a young field. *Arch. Phys. Med. Rehabil.* 2010;91(9 Suppl):S38-43.

252. Ring L, Hofer S, Heuston F, Harris D, O'Boyle CA. Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. *Health Qual. Life Outcomes.* 2005;3:55.

253. Ahmed S, Bourbeau J, Maltais F, Mansour A. The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. *J. Clin. Epidemiol.* 2009;62(11):1165-1172.

254. Mayo NE, Scott SC, Ahmed S. Case management poststroke did not induce response shift: the value of residuals. *J. Clin. Epidemiol.* 2009;62(11):1148-1156.

255. Ramachandran S, Lundy JJ, Coons SJ. Testing the measurement equivalence of paper and touch-screen versions of the EQ-5D visual analog scale (EQ VAS). *Qual. Life Res.* 2008;17(8):1117-1120.

256. Dillman DA, Smyth JD, Christian LM. *Internet, mail, and mixed-mode surveys: the tailored design method.* Hoboken, N.J.: Wiley & Sons; 2009.

257. Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. *Stat. Med.* 1998;17(5-7):653-666.

258. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* Hoboken, NJ: John Wiley & Sons, Inc.; 2002.

259. Keeter S, Kennedy C, Dimock M, Best J, Craighill P. Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey. *Public Opin. Q.* 2006;70(5):759-779.

260. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. *JAMA.* 2012;307(17):1805-1806.

261. Johnson TP, Holbrook AL, Ik Cho Y, Bossarte RM. Nonresponse Error in Injury-Risk Surveys. *Am. J. Prev. Med.* 2006;31(5):427-436.

262. Cull WL, O'Connor KG, Sharp S, Tang S-fS. Response Rates and Response Bias for 50 Surveys of Pediatricians. *Health Serv. Res.* 2005;40(1):213-226.

263. Purdie DM, Dunne MP, Boyle FM, Cook MD, Najman JM. Health and demographic characteristics of respondents in an Australian national sexuality survey: comparison with population norms. *J. Epidemiol. Community Health.* 2002;56(10):748-753.

264. Voigt LF, Koepsell TD, Daling JR. Characteristics of telephone survey respondents according to willingness to participate. *Am. J. Epidemiol.* 2003;157(1):66-73.

265. Fairclough DL. Design and analysis of quality of life studies in clinical trials. Boca Raton: Chapman & Hall/CRC Press; 2002.

266. Little RA. Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc.* 1995;90(431):1112-1121.

267. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for MIXED Models.* Cary, NC: SAS Institute, Inc.; 1996.

268. Hahn EA, Glendenning GA, Sorensen MV, et al. Quality of life in patients with newly diagnosed chronic phase chronic myeloid leukemia on imatinib versus interferon alfa plus low-dose cytarabine: Results from the IRIS Study. *J. Clin. Oncol.* 2003;21(11):2138-2146.

269. Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Stat. Med.* 1998;17(5-7):781-796.

270. Basch EM, Reeve BB, Mitchell SA, et al. Electronic toxicity monitoring and patient-reported outcomes. *Cancer J.* 2011;17(4):231-234.

271. Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients? *Qual. Life Res.* 2007;16(7):1097-1105.

272. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Qual. Life Res.* 2000;9(8):887-900.

273. Deyo RA, Patrick DL. Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Med. Care.* 1989;27(3 Suppl):S254-268.

274. Lipscomb J, Donaldson MS, Arora NK, et al. Cancer outcomes research. *J Natl Cancer Inst Monogr.* 2004(33):178-197.

275. Snyder CF, Aaronson NK, Choucair AK, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual. Life Res.* 2011:S76–S85.

276. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* 2010;63(7):737-745.

277. Revicki DA, Gnanasakthy A, Weinfurt K. Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: the PRO Evidence Dossier. *Qual. Life Res.* 2007;16(4):717-723.

278. Schunemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health Qual. Life Outcomes.* 2006;4:62.

279. Butt Z, Reeve B. Enhancing the Patient's Voice: Standards in the Design and Selection of Patient-Reported Outcomes Measures (PROMs) for Use in Patient-Centered Outcomes Research. 2012; http://www.pcori.org/assets/Enhancing-the-Patients-Voice-Standards-in-the-Design-and-Selection-of-Patient-Reported-Outcomes-Measures-for-Use-in-Patient-Centered-Outcomes-Research.pdf. Accessed June 15, 2012.

280. U. S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. 2009; http://purl.access.gpo.gov/GPO/LPS113413.

281. US Food and Drug Administration. Draft Guidance for industry. Qualification process for drug development tools.
2010; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf. Accessed November 26, 2011.

282. Erickson P, Willke R, Burke L. A concept taxonomy and an instrument hierarchy: tools for establishing and evaluating the conceptual framework of a patient-reported outcome

(PRO) instrument as applied to product labeling claims. *Value Health.* 2009;12(8):1158-1167.

283. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health.* 2007;10 Suppl 2:S125-137.

284. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality of life instruments: attributes and review criteria. *Qual. Life Res.* 2002(11):193-205.

285. Mokkink LB, Terwee CB, Gibbons E, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Med. Res. Methodol.* 2010;10:82.

286. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med. Res. Methodol.* 2011;11(1):152.

287. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med. Res. Methodol.* 2010;10:22.

288. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual. Life Res.* 2010;19(4):539-549.

289. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual. Life Res.* 2011;21(4):651-657.

290. Johnson C, Aaronson N, Blazeby JM, et al. EORTC Quality of Life Group: Guidelines for Developing Questionnaire Modules. 2011; 4th:http://groups.eortc.be/qol/Pdf%20presentations/Guidelines%20for%20Developing%20questionnaire-%20FINAL.pdf. Accessed November 26, 2011.

291. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. 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 PRO Task Force Report. *Value Health.* 2009;12(8):1075-1083.

292. Wild D, Grove A, Martin M, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health.* 2005;8(2):94-104.

293. Magasi S, Ryan G, Revicki D, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. *Qual. Life Res.* 2011.

294. Valderas JM, Ferrer M, Mendivil J, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. *Value Health.* 2008;11(4):700-708.

295. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* 2008;61(2):102-109.

296. PROMIS Validity Standards Committee on behalf of the PROMIS Network of Investigators. The PROMIS® Instrument Development and Psychometric Evaluation Scientific Standards. 2012.

297. Ahmed S, Berzon RA, Revicki DA, et al. The Use of Patient-reported Outcomes (PRO) Within Comparative Effectiveness Research: Implications for Clinical Practice and Health Care Policy. *Med. Care.* 2012;Epub ahead of Print.

298. Kazis LE, Miller DR, Skinner KM, et al. Applications of methodologies of the Veterans Health Study in the VA healthcare system: conclusions and summary. *J. Ambulatory Care Manage.* 2006;29(2):182-188.

299. Kazis LE, Selim A, Rogers W, Ren XS, Lee A, Miller DR. Dissemination of methods and results from the veterans health study: final comments and implications for future monitoring strategies within and outside the veterans healthcare system. *J. Ambulatory Care Manage.* 2006;29(4):310-319.

300. Haffer SC, Bowen SE. Measuring and improving health outcomes in Medicare: the Medicare HOS program. *Health Care Financ. Rev.* 2004;25(4):1-3.

301. National Committee for Quality Assurance, Committee on Performance Measurement. *HEDIS 2006: health plan employer data & information set.* Washington, DC: National Committee for Quality Assurance; 2006.

302. Jordan JE, Osborne RH, Buchbinder R. Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses. *J. Clin. Epidemiol.* 2011;64(4):366-379.

303. Cella D, Nowinski C. Measuring quality of life in chronic illness: The Functional Assessment of Chronic Illness Therapy Measurement System. *Arch. Phys. Med. Rehabil.* 2002;83(Suppl. 2):S10-S17.

304. FDA Center for Drug Evaluation and Research Quality of Life Subcommittee Oncologic Drugs Advisory Committee. Meeting transcript. February 10, 2000. 2000; http://www.fda.gov/ohrms/dockets/ac/00/backgrd/3591b1a.pdf.

305. Shearer D, Morshed S. Common generic measures of health related quality of life in injured patients. *Injury.* 2011;42(3):241-247.

306. Owolabi MO. Which Is More Valid for Stroke Patients: Generic or Stroke-Specific Quality of Life Measures? *Neuroepidemiology.* 2010;34(1):8-12.

307. Bergland A, Thorsen H, Kåresen R. Association between generic and disease-specific quality of life questionnaires and mobility and balance among women with osteoporosis and vertebral fractures. *Aging Clin. Exp. Res.* 2011;23(4):296-303.

308. Rothrock N, Hays R, Spritzer K, Yount SE, Riley W, Cella D. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). *J. Clin. Epidemiol.* 2010;63(11):1195-1204.

309. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. *J. Rheumatol.* 2007;34(6):1426-1431.

310. Donaldson G. Patient-reported outcomes and the mandate of measurement. *Qual. Life Res.* 2008;17(10):1303-1313.

311. Lai JS, Cella D, Choi SW, et al. How Item Banks and Their Application Can Influence Measurement Practice in Rehabilitation Medicine: A PROMIS Fatigue Item Bank Example. *Arch. Phys. Med. Rehabil.* 2011;92(10 Supplement):S20-S27.

312. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J. Clin. Epidemiol.* 2008;61(1):17-33.

313. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J. Chronic Dis.* 1985;38(1):27-36.

314. McClendon DT, Warren JS, Green KM, Burlingame GM, Eggett DL, McClendon RJ. Sensitivity to change of youth treatment outcome measures: a comparison of the CBCL, BASC-2, and Y-OQ. *J. Clin. Psychol.* 2011;67(1):111-125.

315. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual. Life Res.* 2003;12(4):349-362.

316. Beaton DE, van Eerd D, Smith P, et al. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. *J. Clin. Epidemiol.* 2011;64(5):487-496.

317. Andresen EM, Meyers AR. Health-related quality of life outcomes measures. *Arch. Phys. Med. Rehabil.* 2000;81(12 Suppl 2):S30-45.

318. Vermeersch DA, Lambert MJ, Burlingame GM. Outcome Questionnaire: item sensitivity to change. *J. Pers. Assess.* 2000;74(2):242-261.

319. Shikiar R, Willian MK, Okun MM, Thompson CS, Revicki DA. The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study. *Health Qual. Life Outcomes.* 2006;4:71.

320. Schroter S, Lamping DL. Responsiveness of the coronary revascularisation outcome questionnaire compared with the SF-36 and Seattle Angina Questionnaire. *Qual. Life Res.* 2006;15(6):1069-1078.

321. Kaplan RM, Tally S, Hays RD, et al. Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. *J. Clin. Epidemiol.* 2011;64(5):497-506.

322. Bauer S, Lambert MJ, Nielsen SL. Clinical significance methods: a comparison of statistical techniques. *J. Pers. Assess.* 2004;82(1):60-70.

323. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *Jouornal of Clinical Epidemiology.* 2003;56(5):395-407.

324. Brozek JL, Guyatt GH, Schunemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. *Health Qual. Life Outcomes.* 2006;4(69):1-7

325. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control. Clin. Trials.* 1989;10(4):407-415.

326. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual. Life Res.* 1993;2(3):221-226.

327. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J. Pain.* 2008;9(2):105-121.

328. Revicki D, Hays R, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* 2008;61(2):102-109.

329. Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Rev. Pharmacoecon. Outcomes Res.* 2004;4(5):515-523.

330. Guyatt GH. Making sense of quality-of-life data. *Med. Care.* 2000;38(9 Suppl):II175-179.

331. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. *J. Clin. Epidemiol.* 2002;55(9):900-908.

332. Rejas J, Pardo A, Ruiz MA. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *J. Clin. Epidemiol.* 2008;61(4):350-356.

333. Norman G, Sloan J, Wyrwich K. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med. Care.* 2003;41(5):582-592.

334. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann. Intern. Med.* 2006;144(10):742-752.

335. Goldzweig CL, Maglione M, Shekelle PG, Towfigh A. Costs and benefits of health information technology: New trends from the literature. *Health Aff. (Millwood).* 2009;28(2):w282-w293.

336. Wilson EV. *Patient-centered e-health.* Hershey, PA: Medical Information Science Reference; 2009.

337. Harris Interactive, ARiA Marketing. *Healthcare Satisfaction Study.* Rochester, NY: Harris Interactive; 2000.

338. Davis F. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *Int J Man Mach Stud.* 1993;38(3):475-487.

339. United States Department of Health and Human Services, Office of the National Coordinator for Health Information Technology. Meaningful use. 2011; http://healthit.hhs.gov/portal/server.pt?open=512&objID=2996&mode=2. Accessed July, 2011.

340. Estabrooks PA, Boyle M, Emmons KM, et al. Harmonized patient-reported data elements in the electronic health record: supporting meaningful use by primary care action on health behaviors and key psychosocial factors. *J. Am. Med. Inform. Assoc.* 2012;Epub ahead of print.

341. Bitton A, Flier LA, Jha AK. Health information technology in the era of care delivery reform: to what end? *JAMA.* 2012;307(24):2593-2594.

342. Adler-Milstein J, Jha AK. Sharing clinical data electronically: A critical challenge for fixing the health care system. *JAMA.* 2012;307(16):1695-1696.

343. Masys D, Baker D, Butros A, Cowles KE. Giving patients access to their medical records via the internet: the PCASSO experience. *J. Am. Med. Inform. Assoc.* 2002;9(2):181-191.

344. Nelson EC, Hvitfeldt H, Reid RM, et al. *Using Patient-Reported Information to Improve Health Outcomes and Health Care Value: Case Studies from Dartmouth, Karolinska and Group Health.* Lebanon, NH: Dartmouth Institute for Health Policy and Clinical Practice;2012.

345. Davis K, Yount S, Del Ciello K, et al. An innovative symptom monitoring tool for people with advanced lung cancer: A pilot demonstration. *J. Support. Oncol.* 2007;5(8):381-387.

346. Harris WH, Sledge CB. Total hip and total knee replacement (1). *N Engl J Med.* 1990;323(11):725-731.

347. Harris WH, Sledge CB. Total hip and total knee replacement (2). *N Engl J Med.* 1990;323(12):801-807.

348. Liang MH, Cullen KE, Poss R. Primary total hip or knee replacement: evaluation of patients. *Ann. Intern. Med.* 1982;97(5):735-739.

349. Kroll MA, Otis JC, Sculco TP, et al. The relationship of stride characteristics to pain before and after total knee arthroplasty. *Clin. Orthop.* 1989(239):191-195.

350. Ethgen O, Bruyère O, Richy F, Dardennes C, Reginster JY. Health-related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. *J. Bone Joint Surg. Am.* 2004;86-A(5):86.

351. Birrell F, Johnell O, Silman A. Projecting the need for hip replacement over the next three decades: influence of changing demography and threshold for surgery. *Ann. Rheum. Dis.* 1999;58(9):569-572.

352. Rissanen P, Aro S, Sintonen H, Asikainen K, Slätis P, Paavolainen P. Costs and cost-effectiveness in hip and knee replacements. A prospective study. *Int. J. Technol. Assess. Health Care.* 1997;13(4):575-588.

353. Williams MH, Newton JN, Frankel SJ, Braddon F, Barclay E, Gray JAM. Prevalence of Total Hip Replacement: How Much Demand Has Been Met? *J. Epidemiol. Community Health.* 1994;48(2):188-191.

354. Bellamy N. *WOMAC Osteoarthritis Index: user guide IX.* Brisbane: Nicholas Bellamy; 2008.

355. Pua YH, Cowan SM, Wrigley TV, Bennell KL. Discriminant Validity of the Western Ontario and McMaster Universities Osteoarthritis Index Physical Functioning Subscale in Community Samples With Hip Osteoarthritis. *Arch. Phys. Med. Rehabil.* 2009;90(10):1772-1777.

356. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J. Rheumatol.* 1988;15(12):1833-1840.

357. Dunbar MJ, Robertsson O, Ryd L, Lidgren L. Appropriate questionnaires for knee arthroplasty. Results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry. *J. Bone Joint Surg. Br.* 2001;83(3):339-344.

358. McConnell S, Kolopack P, Davis AM. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. *Arthritis Rheum.* 2001;45(5):453-461.

359. Bellamy N, Buchanan WW. A preliminary evaluation of the dimensionality and clinical importance of pain and disability in osteoarthritis of the hip and knee. *Clin. Rheumatol.* 1986;5(2):231-241.

360. Bullens P, van Loon C, de Waal Malefijt M, Laan R, Veth R. Patient satisfaction after total knee arthroplasty. *J. Arthroplasty.* 2001;16(6):740-747.

361. Robertsson O, Dunbar MJ. Patient satisfaction compared with general health and disease-specific questionnaires in knee arthroplasty patients. *J. Arthroplasty.* 2001;16(4):476-482.

362. Davies GM, Watson DJ, Bellamy N. Comparison of the responsiveness and relative effect size of the western Ontario and McMaster Universities Osteoarthritis Index and the short-form Medical Outcomes Study Survey in a randomized, clinical trial of osteoarthritis patients. *Arthritis Care Res.* 1999;12(3):172-179.

363. Bellamy N, Wilson C, Hendrikz J. Population-based normative values for the Western Ontario and McMaster (WOMAC) Osteoarthritis Index: part I. *Semin. Arthritis Rheum.* 2011;41(2):139-148.

364. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann. Rheum. Dis.* 2005;64(1):29-33.

365. Marshall D, Pericak D, Grootendorst P, et al. Validation of a Prediction Model to Estimate Health Utilities Index Mark 3 Utility Scores from WOMAC Index Scores in Patients with Osteoarthritis of the Hip. *Value Health.* 2008;11(3):470-477.

366. Tubach F, Baron G, Falissard B, et al. Using patients' and rheumatologists' opinions to specify a short form of the WOMAC function subscale. *Ann. Rheum. Dis.* 2005;64(1):75-79.

367. Bellamy N, Patel B, Davis T, Dennison S. Electronic data capture using the Womac NRS 3.1 Index (m-Womac): A pilot study of repeated independent remote data capture in OA. *Inflammopharmacology.* 2010;18(3):107-111.

368. Bellamy N, Wilson C, Hendrikz J, et al. Osteoarthritis Index delivered by mobile phone (m-WOMAC) is valid, reliable, and responsive. *J. Clin. Epidemiol.* 2011;64(2):182-190.

369. Theiler R, Bischoff-Ferrari HA, Good M, Bellamy N. Responsiveness of the electronic touch screen WOMAC 3.1 OA Index in a short term clinical trial with rofecoxib. *Osteoarthritis Cartilage.* 2004;12(11):912-916.

370. American College of Rheumatology. Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). 2011; http://www.rheumatology.org/practice/clinical/clinicianresearchers/outcomes-instrumentation/WOMAC.asp. Accessed July 6, 2012.