Patient-Reported Outcomes in Performance Measurement

David Cella, Elizabeth A. Hahn, Sally E. Jensen, Zeeshan Butt, Cindy J. Nowinski, Nan Rothrock, Kathleen N. Lohr
Patient-Reported Outcomes in Performance Measurement

David Cella, Elizabeth A. Hahn,
Sally E. Jensen, Zeeshan Butt, Cindy J. Nowinski,
Nan Rothrock, Kathleen N. Lohr
## Contents

**Introduction**
- The National Quality Forum Project 2
- Concepts and Definitions 3

**Types of Patient-Reported Outcomes**
- Health-Related Quality of Life 8
- Functional Status 9
- Symptoms and Symptom Burden 9
- Health Behaviors 10
- Patient Experience of Care 11

**Method and Mode of Administration, Data Collection, and Analysis**
- Modes and Methods Issues 17
  - Mode: Self-Administration Versus Interviewer Administration 20
  - Method of Administration 21
  - Patient-Reported Outcome Measures in the Clinical Setting 23
  - Scoring: Classical Test Theory Versus Modern Test Theory 26
  - Linking or Cross-Talk Between Different Measures of the Same Construct 27
- Addressing Barriers to Patient-Reported Outcomes Measurement 28
  - Vulnerable Populations 28
  - Literacy 29
  - Language and Culture 30
  - Functional Abilities 31
  - Response Shift, Adaptation, and Other Challenges to Detecting True Change 33
- Implications of the Different Methods and Modes for Response Rate, Reliability, and Validity 34
  - Implementing Data Collection Methods 34
  - Using Different Method or Mode Than the One Originally Validated 34
  - Using Multiple Methods and Modes 35
  - Accounting for the Impact of Nonresponders 35
- Selection of Patient-Level PROMs 36
  - Patient-Centered Outcomes Research 36
  - Attributes of PROMs 37

*(continued)*
Important Differences in PROM Attributes 42
Instrument Length 42
Implications of PRO Data for Action 43
History of Successful Use of PROMs 43
Documentation of Particular Attributes of PROMs 44
PROM Characteristics for Consideration 45
Generic Versus Condition-Specific Measures 45
Measurement Precision 47
Sensitivity to Change, or Responsiveness 47
Minimally Important Differences 49
Essential Conditions to Integrate PROMs Into the Electronic Health Record 51
General Considerations for Health Information Technology 51
Examples of PROMs in Electronic Health Record Applications 53
Selection of PROMs That Meet Recommended Characteristics for Use in Performance Measures 55
Conclusions 57
References 59
Acknowledgments 93
About the Authors 95

Tables
Table 1. Definitions and key concepts for patient-reported outcomes and measures 4
Table 2. Main characteristics of patient-reported outcomes 7
Table 3. PRO methods: characteristics, strengths, and limitations 15
Table 4. Primary criteria for evaluating and selecting patient-reported outcome measures (PROMs) for use in performance measurement 38

Figure
Figure 1. Types of respondent sources of data and modes and methods of administration 18
Introduction

The increasing integration of health care delivery systems provides an opportunity to manage entire episodes of care in a patient-focused manner and to assess the impact of care on patient outcomes, including patient-reported outcomes (PROs). The National Institutes of Health (NIH) PROMIS initiative (for Patient Reported Outcomes Measurement Information System) describes PROs as direct feedback from patients “on their feelings or what they are able to do as they are dealing with chronic diseases or conditions.” They reflect the health status or health-related quality of life circumstances of patients (broadly defined). Such information is reported by individuals themselves or, in some cases, by proxy respondents such as parents for young children or close relatives of persons unable to report for themselves.

PRO information is widely gathered in clinical and health services research; it is increasingly collected and used in clinical practice settings as well. PROs are relevant for many activities: helping patients and their clinicians make informed decisions about health care, monitoring the progress of care, setting policies for coverage and reimbursement of health services, improving the quality of health care services, and tracking or reporting on the performance of health care delivery organizations.

These last two activities are gaining increasing attention in the US health care system. The nation is engaging in more efforts to expand health care coverage to many millions of citizens through the Patient Protection and Affordable Care Act (ACA). Many organizations are working to ensure higher value of health care through enhanced attention to measuring and improving quality of care and patient outcomes. For example, the Patient-Centered Outcomes Research Institute (PCORI), established by the ACA, is actively pursuing ways to increase the use of PROs for clinical care, research, and performance assessment.
The National Quality Forum Project

The National Quality Forum (NQF), a national organization that has been deeply involved in moving the quality-improvement agenda forward for many years (http://www.qualityforum.org), endorses and promulgates quality-of-care and performance measures that various provider groups, regulatory agencies, payers and insurers, and others can use for accountability and quality improvement activities. In 2012 and 2013, the organization began an initiative to find PROs that might be added to its extensive collection of performance measures. Its National Voluntary Consensus Standards for Patient Outcomes: A Consensus Report defined outcomes as being important because they “reflect the reason that an individual seeks healthcare services.”1 The individual patient’s voice in many performance measures, however, has largely been missing. Few ways to assess performance are available at the organizational level, even though patients are often the best able to report on the experiences and results of their individual care.

To fill that gap, NQF convened an expert panel at two public meetings and, as background for its deliberations, commissioned two authoritative background papers. This monograph is a revised and updated version of the first of these two papers, which provided the background on issues about selecting PROs for use in a variety of applications pertinent to the NQF mission and activities. The second paper, Patient-Reported Outcomes in Performance Measurement,9 dealt with issues relating to processes for endorsing performance measures that reflect the end results (ultimate outcomes) of health care. Its primary focus was on accountable health care organizations.

This monograph applies the conceptual and organizational frameworks that NQF has pioneered in the past decade or so. NQF distinguishes PROs, patient-reported outcome measures (or PROMs), and patient-reported outcome performance measures (or PRO-PMs). NQF endorses PRO-PMs through transparent and consensus-based procedures. This monograph addresses the PROs that are likely to be used to inform PRO measures (PROMs) that would underpin scientifically acceptable and feasible performance measures. We do not address issues with identifying, evaluating, or endorsing PRO-PMs for health care organizations or clinicians.
To accomplish our assigned objective, we completed a comprehensive review of the published peer-reviewed literature as well as published documents (e.g., book chapters, position statements; guidance from the US Food and Drug Administration) about standard measurement theory. Based on our findings in the published literature, we created a comprehensive annotated outline reflecting the methodological considerations most important to address. We then revised it in consultation with representatives from NQF to ensure that the paper accomplished the NQF’s high-priority objectives.

Following the initial drafting of the manuscript, our group of authors at Northwestern University worked in conjunction with Kathleen Lohr, RTI International, who comprehensively reviewed and revised the manuscript draft. We submitted the revised manuscript draft for review and comment at the NQF Patient Reported Outcomes Workshop 1 in July 2012. A final version of the manuscript, which incorporated revisions based on feedback from the workshop, was submitted to NQF in September 2012. That manuscript includes numerous authoritative citations to research from this field up through that date. Dr. Lohr and the other authors revised the manuscript further to meet the requirements of an RTI Press monograph and to update some of the citations; the result is this monograph.

**Concepts and Definitions**

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 (Table 1). There are many available PRO measurement tools, which we refer to here as patient-reported outcome measures, or PROMs. By using direct, unfiltered inquiry, PROMs measure what patients are able to do and how they feel. They reflect the direct voice of the patient, as perceived by the patient.
Table 1. Definitions and key concepts for patient-reported outcomes and measures

<table>
<thead>
<tr>
<th>Key Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>A person who is receiving health care services or using long-term health care support services.</td>
</tr>
<tr>
<td>Patient-reported outcome (PRO)</td>
<td>Any information on the outcomes of health care obtained directly from patients without modification by clinicians or other health care professionals. For purposes of this monograph, we use this term broadly to include any patient input, whether or not it is standardized or gathered with a structured questionnaire.</td>
</tr>
<tr>
<td>Patient-reported outcome measure (PROM)</td>
<td>Any standardized or structured questionnaire regarding the status of a patient’s health condition, health behavior, or experience with health care that comes directly from the patient (i.e., a PRO). The use of a structured, standardized tool such as a PROM will yield quantitative data that enables comparison of patient groups or providers. One example of a PROM is the nine-item Patient Health Questionnaire (PHQ-9).</td>
</tr>
<tr>
<td>Performance measure</td>
<td>Numeric quantification of health care quality for a designated accountable health care entity, such as a hospital, health plan, nursing home, clinician, etc.</td>
</tr>
<tr>
<td>PRO-based performance measure (PRO-PM)</td>
<td>A performance measure that is based on patient-reported outcomes assessed through data often collected through a PROM and then aggregated for an accountable health care entity. One example is the percentage of patients in an accountable care organization with an improved depression score as measured by a standardized tool such as the PHQ-9.</td>
</tr>
<tr>
<td>e-health</td>
<td>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.</td>
</tr>
<tr>
<td>Patient-centered outcomes research (PCOR)</td>
<td>Integration of patient perspectives and experiences with clinical and biological data collected from the patient to evaluate the safety and efficacy of an intervention (<a href="http://www.pcori.org">www.pcori.org</a>).</td>
</tr>
<tr>
<td>Reliability</td>
<td>The extent to which a scale or measure yields reproducible and consistent results. Reliability of data elements refers to repeatability and reproducibility of the data elements for the same population in the same time period. Reliability of the measure score refers to the proportion of variation in the performance scores attributable to systematic differences across the measured entities (or signal) in relation to random error (or noise).</td>
</tr>
<tr>
<td>Validity</td>
<td>The extent to which an instrument measures what it is intended to measure and can be useful for its intended purpose. Validity of instruments can be assessed in numerous ways, often in comparison with an authoritative source (such as a similar validated instrument). Validity of measure scores can refer to the correctness of conclusions that users might draw from a reliable and valid instrument (as, for instance, that a better score on a quality measure reflects higher quality of health care).</td>
</tr>
</tbody>
</table>
Using Patient-Reported Outcome Measures (PROMs)

A large literature supports the use of PROMs and provides cogent evidence suggesting that clinicians are limited in accurately estimating outcomes for patients.\textsuperscript{14–18} PROMs 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 PROMs have been used for clinical research purposes and to guide clinical care. Many have been evaluated and catalogued by the NIH PROMIS network and made available through the PROMIS Assessment Center (www.nihpromis.org/software/assessmentcenter).

Two major challenges to using PROMs for purposes of accountability and performance improvement must be addressed. First, they have not yet 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 organizations and systems.

Many in the health sector are showing increasing interest in moving toward use of PROMs for these clinical, quality improvement, and accountability applications. Foundational work still needs to address methodological and data challenges. Efforts in the early 2010s focused on developing and testing mechanisms for collecting patient-reported data. A crucial element of this is considering methodological issues in some depth. Among the more difficult problems are collecting PRO data in the clinical environment and aggregating data to assess organization- and clinician-level performance.

In the remainder of this monograph, we address the major methodological issues related to the selection, administration, and use of PROMs for individual patients in clinical practice settings. We highlight best practices in identifying and using PROMs in performance measures. Given such information, those concerned with identifying and choosing appropriate PROMs as candidate measures for use in performance assessment and related applications can move ahead in this arena.
Types of Patient-Reported Outcomes

PROMs can be used to assess a wide variety of health-relevant concepts. Of particular salience for quality and performance measurement efforts are the following five categories: 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 PROMs. In the table, we highlight only key advantages or drawbacks for each PRO category. In the subsections that follow, we focus on core components or attributes of the specific category in question of particular relevance for measurement (including efficient performance measurement). Consequently, the information for any given PRO category may differ from that for other categories.

Table 2. Main characteristics of patient-reported outcomes

<table>
<thead>
<tr>
<th>PRO Category</th>
<th>Main Characteristics</th>
<th>Main Strengths</th>
<th>Main Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health-related quality of life (HRQL)</td>
<td>• Is multidimensional</td>
<td>• Yields a global summary of well-being</td>
<td>• May not be considered a sufficiently specific construct</td>
</tr>
<tr>
<td></td>
<td>• Can be generic or condition-specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional status</td>
<td>• Reflects ability to perform specific activities</td>
<td>• Can be used in addition to performance-based measures of function</td>
<td>• May reflect variations in self-reported capability and actual performance of activities</td>
</tr>
<tr>
<td>Symptoms and symptom burden</td>
<td>• Are specific to type of symptom of interest</td>
<td>• Are best assessed through self-report</td>
<td>• May fail to capture general, global aspects of well-being considered important to patients</td>
</tr>
<tr>
<td></td>
<td>• May identify symptoms not otherwise captured by medical workup</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Main characteristics of patient-reported outcomes (continued)

<table>
<thead>
<tr>
<th>PRO Category</th>
<th>Main Characteristics</th>
<th>Main Strengths</th>
<th>Main Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health behaviors</td>
<td>• Are specific to type of behavior&lt;br&gt;• Typically measure frequency of behavior</td>
<td>• Target specific behavior categories</td>
<td>• Validity may be affected by social desirability&lt;br&gt;• May produce potential patient discomfort in reporting socially undesirable behaviors</td>
</tr>
<tr>
<td>Patient experience</td>
<td>• Concerns satisfaction with health care delivery, treatment recommendations, and medications (or other therapies)&lt;br&gt;• Reflects actual experiences with health care services&lt;br&gt;• Fosters patient activation</td>
<td>• Is an essential component of patient-centered care&lt;br&gt;• Is valued by patients, families, and policy makers&lt;br&gt;• Relates to treatment adherence&lt;br&gt;• Relates to health behaviors and health outcomes</td>
<td>• May be a complex, multidimensional construct&lt;br&gt;• Requires confidentiality to ensure patient comfort in disclosing negative experiences&lt;br&gt;• Does not provide sufficient evidence that activation enhances health care decision making</td>
</tr>
</tbody>
</table>

Health-Related Quality of Life

One class of PRO measures health-related quality of life (HRQL). HRQL is a multidimensional\(^\text{19}\) construct encompassing physical, social, and emotional well-being associated with illness and its treatment.\(^\text{20}\) Different types of HRQL measures\(^\text{21,22}\) are useful for different purposes.\(^\text{23}\) 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.\(^\text{24-27}\) This type of PROM is useful in assessing individuals both with and without a health condition. Such data allow 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.\(^\text{28}\) Experts can use scores from these types of measures to calculate quality-adjusted life years or compare information to population norms.
Many PROMs are intended for use in populations with chronic illnesses.\textsuperscript{29–31} Over the past 8 years, the PROMIS network has developed a considerable number of PROMs in physical, mental, and social health for adults and infants, children, and adolescents with chronic conditions.\textsuperscript{32,33} Neuro-QOL is another measurement effort focused on capturing important areas of functioning and well-being in neurologic diseases.\textsuperscript{34} These measurement efforts do not reference a specific disease in the items; thus, they permit comparisons across conditions. Other PROMs are targeted on a specific disease (e.g., spinal cord injury) or treatment (e.g., chemotherapy).\textsuperscript{35,36} Often these instruments are developed so that investigators can demonstrate responsiveness to treatment in a clinical trial rather than compare data against population norms or information on other conditions.\textsuperscript{37} Condition-specific PROMs often provide additional, complementary information about a patient’s HRQL.\textsuperscript{30,38–40}

**Functional Status**

Another type of PROM 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.\textsuperscript{41} Examples of functional status include physical function, cognitive function, and sexual function. As with HRQL instruments, a large number of functional status measures exist, but they vary widely in quality.\textsuperscript{42} 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), whereas others may be appropriate for use across chronic conditions.\textsuperscript{43–49}

**Symptoms and Symptom Burden**

Symptoms such as fatigue and pain intensity are key domains for PROMs. Symptoms are typically negative, and their presence and intensity are best assessed through patient report.\textsuperscript{50} Scales 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 PROMs. Symptom burden captures the combination of both symptom severity and impact experienced with a specific disease or treatment.\textsuperscript{50}

Common symptom and symptom burden measures include the Functional Assessment of Chronic Illness Therapy—Fatigue scale, which is not targeted on any one condition. By contrast, disease-focused symptom indexes include the symptom indexes for various cancer types set out by the National Comprehensive Cancer Network and a dyspnea-specific instrument for
chronic obstructive pulmonary disease.51,52 PROMIS investigators developed the PROMIS Pain Interference measure, which quantifies the impact of pain on functioning.53

Health Behaviors

Yet another category of PROMs 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 PROMs serves several important clinical purposes. Clinicians can use it to monitor risk behaviors with potentially deleterious health consequences. This information enables practitioners to identify areas for risk reduction and health promotion interventions among their patients. Health behavior PROMs can also be used to assess patients’ response to health promotion interventions and to monitor health behaviors over time.

Health risk assessments (HRAs) illustrate how health behavior PROMs can be incorporated into health promotion and disease prevention programs. Defined by the US Centers for Disease Control and Prevention (CDC) as tools to measure individual health, HRAs may consist of clinical examination or laboratory test results as well as health behavior PROMs.54 A recent report from the US Agency for Healthcare Research and Quality (AHRQ) identified three key components in the process of implementing HRAs in health promotion: (1) patient self-reported information to identify risk factors for disease, (2) individualized health-specific feedback to patients based upon the information they reported, and (3) at least one health promotion recommendation or intervention.55

Although HRAs have been implemented in community settings, universities, and health maintenance organizations, they have been most commonly implemented in workplace settings.55 An extensive review of HRA program outcomes concluded that, in many cases, implementing HRA programs improved health behaviors and intermediate health outcomes (e.g., blood pressure); however, the evidence did not demonstrate whether using HRAs affected disease incidence or health outcomes over the medium to long term.55

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
Types of Patient-Reported Outcomes

category. Two examples of health behavior PROMs measuring multiple risk factors that the National Committee for Quality Assurance has certified are the Personal Wellness Profile\(^5\) and the Insight Health Risk Appraisal Survey.\(^5\)

In addition, several large-scale health behavior assessment systems provide additional context for the use of general health behavior PROMs. The Behavioral Risk Factor Surveillance System (BRFSS), created in 1984 by the CDC as a state-based system, uses a standardized questionnaire to measure health risk and health promotion behaviors. These include health awareness, tobacco use, consumption of fruits and vegetables, physical activity, seatbelt use, immunization, and alcohol consumption.\(^5\) The National Health and Nutrition Examination Survey (NHANES) constitutes another large-scale implementation of health behavior PROMs. Established by the CDC in the 1960s, NHANES includes health behavior surveys in addition to clinical examinations to assess health status at the population level.\(^5\)

The health behavior survey portion of NHANES assesses a wide range of health risk and health promotion behaviors, including smoking, drug use, alcohol use, sexual practices, physical activity, dietary intake, and reproductive health practices.\(^5\) Health behavior PROMs can also assess risk factors associated with specific diseases (e.g., smoking) or those related to specific behavioral categories (e.g., physical activity, seatbelt use, food consumption). The health risk survey, an interactive computer-based survey assessing alcohol consumption and smoking,\(^6\) is one example. Another is the CAGE-Adapted to Include Drugs (CAGE-AID) questionnaire, a self-reported screening measure of substance use disorder among treatment-seeking adolescents. Its name derives from its four main questions (Cutting down, being Annoyed if people criticize drinking, feeling Guilty about drinking, and needing an Eye-opener).\(^6\)

A subset of health behavior PROMs assesses health-promoting behaviors. Examples of such PROM instruments include “Starting the conversation,” a brief measure of dietary intake;\(^6\) “Exercise as the fifth vital sign,” a brief measure of physical activity;\(^6\) School Health Action, Planning and Evaluation System (SHAPES), a school-based self-report physical activity measure;\(^6\) and the Morisky Medication Adherence Scale (8-item).\(^6\)

**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 (now known as the National Academy 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. Measurement of patient ratings is a complex concept that is related to perceived needs, expectations of care, and experience of care. 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 tailor treatments based on informed decisions that their patients might make based on those preferences. In fact, improving decision quality is one critically important step that the nation can take to improve the quality (processes and outcomes) of health care and thus enhance value for health care expenditures. For this reason, patients’ ratings of their experiences with care not only 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.”

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. Shikiar and Rentz proposed a three-level hierarchy of satisfaction: (1) satisfaction with health care delivery, including issues of accessibility, clinician-patient communication, and quality of facilities; (2) satisfaction with the treatment regimen, including medication, dietary and exercise recommendations, and similar elements of therapies; 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. It is an important indicator of future adherence to treatment. Satisfaction has a long history of measurement, and numerous instruments are available.

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. Hibbard and colleagues developed a 13-item scale, the Patient Activation
Measure (PAM),\textsuperscript{103,104} which demonstrated favorable psychometric properties in several cross-sectional and some longitudinal studies.\textsuperscript{101} Although appreciation of the benefits of activated patients is increasing,\textsuperscript{105} commensurate support is lacking to help patients become more activated with respect to their health care decision making.\textsuperscript{104} Although research supports the claim that improvements in patient activation are associated with improvements in self-reported health behaviors,\textsuperscript{101,105} additional research is necessary to better understand both these relationships and their relevance to actual behavior. Patient activation, as measured by the PAM or otherwise, may be a useful moderator or mediator of PROs that will in turn contribute to performance measurement.

An important contemporary focus is on measuring patient reports of their actual experiences with health care services.\textsuperscript{106} Reports about care are often regarded as more specific, actionable, understandable, and objective than general ratings alone.\textsuperscript{107,108} The Consumer Assessment of Healthcare Providers and Systems (CAHPS) program is a multiyear AHRQ initiative to support and promote the assessment of consumers’ experiences with health care. 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.
Method and Mode of Administration, Data Collection, and Analysis

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 PROMs. Numerous issues complicate scoring and analyzing PROM response data. We first describe these methods issues (Table 3)—sources of reports, modes of administration, methods of administration, settings, and scoring—and then discuss barriers.

As with the earlier descriptions of core PRO categories such as health-related quality of life, we highlight in this section the critical issues for measurement methods—i.e., the advantages or drawbacks that users would most need to take into account. This information reflects standard measurement theory (classical or contemporary) and is based on decades of published research and theoretical papers and inputs from experts involved with projects such as PROMIS.

Table 3. PRO methods: characteristics, strengths, and limitations

<table>
<thead>
<tr>
<th>Methodological Issue</th>
<th>Main Characteristics</th>
<th>Main Strengths</th>
<th>Main Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of report</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>Person responds about himself or herself</td>
<td>• Expert on own experience</td>
<td>• Not always possible to assess directly, e.g., because of cognitive or communication deficits or age/developmental level</td>
</tr>
<tr>
<td>Proxy</td>
<td>Person responds about someone else</td>
<td>• Useful when target of assessment is unable to respond • Can provide complementary information</td>
<td>• May not accurately represent subjective or other experiences</td>
</tr>
<tr>
<td>Methodological Issue</td>
<td>Main Characteristics</td>
<td>Main Strengths</td>
<td>Main Limitations</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Mode of administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>Person self-administers PROM and records the responses</td>
<td>• Cost-effective</td>
<td>• Potential for missing data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• May yield more participant disclosure</td>
<td>• Simple survey design (e.g., minimal skip patterns)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Proceed at one’s own pace</td>
<td></td>
</tr>
<tr>
<td>Interviewer</td>
<td>Interviewer reads questions aloud and records the responses</td>
<td>• More complex survey design (e.g., skip patterns)</td>
<td>• Interviewer costs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Useful for respondents with reading, writing, or vision difficulties</td>
<td>• Potential for bias (interviewer bias, social desirability bias, acquiescent response sets)</td>
</tr>
<tr>
<td><strong>Method of administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paper-and-pencil</td>
<td>Patient self-administers PROM using paper and a writing utensil</td>
<td>• Cost-effective</td>
<td>• Prone to data entry errors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Data entry, scoring require more time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Less amenable to incorporation within EHR</td>
</tr>
<tr>
<td>Electronic</td>
<td>Patient self-administers PROM using computer- or telephone-based platform</td>
<td>• Interactive</td>
<td>• Cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Practical</td>
<td>• Potential discomfort with technology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Increased comfort for socially undesirable behaviors</td>
<td>• Accessibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Minimizes data entry errors</td>
<td>• Measurement equivalence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Immediate scoring, feedback</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Amenable to incorporation within EHR</td>
<td></td>
</tr>
<tr>
<td><strong>Setting of administration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic</td>
<td>Patient completes PROMs when he or she arrives to clinic appointments</td>
<td>• Real-time assessment of outcomes</td>
<td>• Impact on clinic flow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Feasibility with use of electronic methods of administration</td>
<td>• Interruptions resulting in missing data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Patient anxiety</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Staff burden</td>
</tr>
</tbody>
</table>

(continued)
Method and Mode of Administration, Data Collection, and Analysis

Table 3. PRO methods: characteristics, strengths, and limitations (continued)

<table>
<thead>
<tr>
<th>Methodological Issue</th>
<th>Main Characteristics</th>
<th>Main Strengths</th>
<th>Main Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Setting of administration (continued)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| **Home** | Patient completes PROMs at home before or between clinic visits | • Minimizes impact on clinic flow  
• Minimizes staff burden | • Accessibility  
• Health information privacy  
• Data security  
• Patient safety |
| **Other** | Patients complete PROMs at other types of settings (e.g., skilled nursing, rehabilitation) | • Feasibility with electronic methods of administration | • Cognitive capacity and potential need for proxy |

<table>
<thead>
<tr>
<th>Scoring</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classical test theory</strong></td>
<td>Raw scores</td>
<td>• Easy to implement and understand</td>
<td>• All items must be administered</td>
</tr>
</tbody>
</table>
| **Modern test theory** | Probabilistic approach | • Enables CAT (tailored questions)  
• Shorter questionnaires with more precision | • Difficult to implement and understand |

CAT, computer-assisted testing; EHR, electronic health record; PRO, patient-reported outcome; PROM, patient-reported outcome measure.

Source: Data are from Data Collection Methods. Quality of Life and Pharmacoeconomics in Clinical Trials\textsuperscript{130} and Psychological Aspects of Health-Related Quality of Life Measurement: Tests and Scales. Quality of Life and Pharmacoeconomics in Clinical Trials.\textsuperscript{131}

### Modes and Methods Issues

Administering PRO instruments (PROMs) requires users to make decisions about three aspects of data collection (Figure 1):

- **Data source**—i.e., the source of the PRO (the patient or, in some cases, a proxy or other reporter)
- **Mode** by which information was recorded—i.e., self-administered or interviewer-administered
- **Method** used to capture the information (such as paper-and-pencil questionnaire or telephone- or computer-assisted technologies).

Each of these aspects is described below. These three aspects can be combined in various ways. For example, a patient might use the telephone to self-administer a PROM, or an interviewer might use a computer to read questions and record answers.
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 PROMs.\textsuperscript{110} This is especially likely to occur with the elderly and with people of any age who have severe disease or suffer from neurological disorders. Children’s participation can be limited by these same factors plus issues specific to their age and developmental level.\textsuperscript{110–112}

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.

**Proxy Report as a Substitute for Self-Report**

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. It not only allows inclusion of a broader and more representative range of patients in the entire measurement effort, but it can also help minimize missing data and increase the feasibility of longitudinal assessment.
The usefulness of proxy responses as substitutes 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, analysts usually compare proxy responses with patient responses. This is a reasonable approach, when proxy responses are being used to replace patient responses.

Agreement between the proxy and patient 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. 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 literatures suggest 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 emotional well-being. Using continuous rather than dichotomous ratings improves agreement. Extent of disagreement increases with increasing age of adolescents and as the severity of patient illness, cognitive impairment, or disability rises. Type of proxy (e.g., parent versus caregiver) and proxy characteristics such as age, education, and level of stress may also affect agreement. 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. Patterns of disagreement for child- versus proxy-reported outcomes are inconsistent. Even when self- and proxy reports disagree for either children or adults, differences tend to be small.

Proxy Report as a Complement to Self-Report

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 an external or “other” rating for the sake of clarity. An important consideration is that the measure make clear which perspective is desired.
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 PROMs but fail to recognize the extent of their impaired well-being and physical role functioning. In such cases, a next-of-kin caregiver such as a spouse could provide an external assessment that indicates that the patient has some degree of problems in functioning, such as getting the groceries from car to kitchen or being comfortable in a social setting. In these circumstances, external (proxy) respondents can clearly introduce clinically important information.

**Mode: Self-Administration Versus Interviewer Administration**

Self-administration of PROMs 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 barrier.

**Main Advantages and Disadvantages of Different Modes of Administration**

Table 3 summarized the principal benefits and drawbacks of different modes of administration, based on authoritative sources. Self-administered instruments are more cost-effective from a staffing perspective, and they may yield more patient disclosure, especially when collecting sensitive information. 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.\textsuperscript{133}

**Additional Concerns About Sources of Bias**

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 or disagree with statements regardless of their content).\textsuperscript{134,135}

Legitimate concerns arise about the potential biasing effects of mode of administration on data quality and interpretation.\textsuperscript{136} 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.\textsuperscript{121–124} For example, some studies have reported more favorable reports of well-being on self-administered questionnaires,\textsuperscript{137} whereas others have found the opposite effect.\textsuperscript{138–140} Still other studies reported mixed results\textsuperscript{141} or found no important differences attributable to mode of administration after adjusting for other factors.\textsuperscript{130,142,143} Fortunately, many types of error and bias can be overcome by appropriate selection and training of interviewers.

**Method of Administration**

Advances in technology have changed the face of PROM assessment, increasing the number of administration options available. Multiple methods of self-administration currently exist, and the different methods may have different effects on the quality of the data.\textsuperscript{136} Although diverse administration methods provide more options for researchers and clinicians, they require different skills and resources of people being asked to respond to the questionnaire. This means that the choice of method of administration may pose differing levels of respondent burden.\textsuperscript{136}

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.\textsuperscript{136} Thus, when users are deciding on one (or more) appropriate methods of administration for a given PROM, they must give these factors due consideration.
Historically, paper-and-pencil administration served as the primary method of PROM assessment. Many PROMs were originally developed with the intention of paper-based administration, but they may be (and typically are) amenable to an electronic-based administration.\textsuperscript{149} Paper-and-pencil remains a widely used PROM administration method, with its primary advantage being cost-effectiveness in situations in which users face 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 scanning of paper-and-pencil PROMs, 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 are important.

Advances in technology and the increasingly widespread availability of electronic resources have provided several alternatives to paper-and-pencil administration. Improved telephone technology has enabled the use of interactive voice response to administer PROMs. Interactive voice response involves a computer audio-recording of PROM questions administered via telephone to which people indicate their responses by selecting the appropriate key.\textsuperscript{136,149}

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/la Pantalla Parlanchina, audiovisual computer-assisted self-interviewing).\textsuperscript{136,149–151} Newer mobile forms of technology such as tablet computers and smartphones also offer promise as methods of PROM administration.

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 PROM administration has been shown to be practical, acceptable, and cost-effective.\textsuperscript{60} Electronic methods may also provide people with increased comfort when responding to questions about socially undesirable behaviors.\textsuperscript{152}

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 PROM administration highlights the importance of measurement equivalence across methods.\textsuperscript{149} Measurement equivalence is determined by comparing the psychometric properties of data obtained via paper-based administration and data collected through electronic administration.\textsuperscript{149} It can be assessed via cognitive testing, usability testing, equivalence testing, or psychometric testing (or various combinations of these techniques).\textsuperscript{149} A growing body of research documents the equivalence of electronic and paper-and-pencil administration of PROMs.\textsuperscript{153–155} These findings support the viability of electronic PROM 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 PROMs, physical transfer of the PROM from patient to provider, as well as the physical existence of the completed PROM, may pose risks to the privacy and confidentiality of patients’ responses. Privacy also emerges as a concern about electronic 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 PROM assessment.

**Patient-Reported Outcome Measures in the Clinical Setting**

Collecting PRO data as part of clinical care has become common.\textsuperscript{2,156–158} Facilitating introduction of these PROMs into clinical practice and decision
making promises many benefits. Advocates for using PROMs in clinical care propose that the results assist clinical providers in managing their patients’ care,\textsuperscript{159} enhance the efficiency of clinical practice,\textsuperscript{155,160} improve patient-provider communication,\textsuperscript{155,160–162} identify patient needs in a timely manner,\textsuperscript{155,163} and facilitate patient-centered care.\textsuperscript{155} Other findings, however, suggest regional variation in perceived health and no positive effect of feedback via PROMs on care, even when combined with guideline-recommended interventions.\textsuperscript{164–166} As PROMs 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 PROMs in the clinic setting.\textsuperscript{155,160,162,163,167–170} When selecting PROMs for administration in clinical practice, users need to consider the efficiency of PROM administration, scoring, and interpretation. These factors are especially important because of the time-sensitive nature of the clinic workflow.\textsuperscript{155,167} In addition, acceptability of both the PROM and the data collection process for both patients and clinic staff is essential.\textsuperscript{155,167,171}

Historically, several barriers have impeded widespread implementation of PRO data collection in clinical settings of all sorts, but especially for smaller or private practices. Many drawbacks are associated with paper-and-pencil administration of PROMs. One such barrier involves concerns about the potential disruption to the clinical workflow if patients are asked to complete PROMs.\textsuperscript{159} In addition, staff burden and clinician disengagement may hamper obtaining PRO data in clinical settings.\textsuperscript{159}

Fortunately, technology advances, and the increased opportunities for methods of PROM administration that they afford, may help to overcome some barriers to PRO data collection in clinical practices and settings.\textsuperscript{167} For example, research supports the feasibility of using tablet computers\textsuperscript{155,163} and touchscreen computers for these purposes.\textsuperscript{150,151,162,167,168} Employing computers to administer PROMs 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 PROMs, which could potentially result in missing data.\textsuperscript{159} Another potential barrier involves the possibility that patients may experience anxiety in completing PROMs in the clinical settings before their appointments.\textsuperscript{159} Similarly, a possible lack of privacy when completing PROMs in waiting rooms or similar circumstances poses another potential obstacle.
to adequate PROM administration. Many of these concerns can be addressed by incorporating PROMs into the clinical workflow. This may also enhance completion rates. Both patients and providers will then be more likely to see this effort as integral to patient care.

Completing PROMs from home before or between medical appointments has been proposed as one strategy to overcoming the problems outlined above.\textsuperscript{159,172,173} Both web-based PROM administration and interactive voice response constitute possible methods for at-home PRO data collection.\textsuperscript{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.\textsuperscript{159,172} First, for patients to be able to complete PROMs at home, they must have access to the type of technology by which the PROM is administered (e.g., Internet). Second, patients must find completing PROMs at home acceptable. Third, users should have a plan in place to address situations in which home-based PROM responses suggest critical or acute problems. This may pose a logistical challenge in comparison with PROMs 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 PROMs is the availability of secure data collection platforms.\textsuperscript{159,174} 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 outcomes that patients report, rather than meeting directly with patients to discuss questions or problems that their patients raise through answers to the PROMs.\textsuperscript{159,174}

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 issues in administering PROMs in these settings. However, handheld technology may offer a means of facilitating collection of PRO data in the rehabilitation setting following orthopedic surgery.\textsuperscript{175}

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 PROMs. 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

Many PROMs involve the measurement of latent (not directly observable) variables; examples might include symptoms (not signs) of gastrointestinal disease or pain. The only way to estimate a person’s level on a particular attribute is by asking questions that represent the attribute in question. Most PROMs 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 either are caused by or are manifestations of the underlying latent attribute, and the items are expected to correlate with one another.176–179

In some other kinds of multi-item measures, the items may cluster together but would not be expected to correlate. A common example of this latter measure is a comorbidity index comprising various health conditions, e.g., diabetes, asthma, and heart disease. Another example might be 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 reflecting underlying constructs with items expected to correlate with one another.

Scoring is based on classical test theory (raw scores) or modern test theory (item response theory [IRT]).180–189 Multiple items are preferred because a response to a single item provides only limited information to distinguish among individuals.190 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.190–192

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.13,190,193–196 This approach requires all items on a particular PROM to be used in every situation for it to be considered valid. Hence, the instrument is test-dependent.194,196–198

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.13,190,196–201 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.\textsuperscript{13}

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.\textsuperscript{13,196,197,202} 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.\textsuperscript{203,204} Typically the metric has been emphasized more than the measure. *Equating* is a technique to convert the system of units of one measure to that of another. Analysts have successfully used this process of deriving equivalent scores 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.\textsuperscript{203} 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 PROMs across studies. Two measures may be considered linked if they produce scores that match the first two moments of their distributions (i.e., mean and standard deviation 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 Patient-Reported Outcomes Measurement

Users need to address yet other barriers to PRO measurement. These include administering PROMs 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 nonresponders to items and questionnaires. In discussing these issues below, we also note best practices and recommendations for addressing them.

Vulnerable Populations

Recognition is growing that some population subgroups are particularly vulnerable to receiving suboptimal health care and to failing to achieve health outcomes equivalent to those experienced by the general population.\textsuperscript{205-207} 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.\textsuperscript{205} 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.\textsuperscript{150,206} It also includes people residing in areas with health professional shortages.\textsuperscript{168}

Administration of PROMs 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, and 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.206,208

**Literacy**

Low literacy is a widespread but neglected problem in the United States. The 1992 National Adult Literacy Survey (NALS)209 and the 2003 National Assessment of Adult Literacy (NAAL)210 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 US population were estimated to be nonliterate in English. Generally speaking, *health* literacy problems complicate matters of both health care delivery and PRO measurement.211,212

*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.”213 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.214,215 Limited health literacy is widespread214,216 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.211,214,217–219 Health literacy may be more limiting than functional literacy because of the unfamiliar context and vocabulary of the health care system.212,214,220

Contributing to poor understanding of the importance of literacy skills is the fact that low literacy is often underreported. The NALS reported that 66 percent to 75 percent of adults in the lowest reading level and 93 percent to 97 percent in the second-lowest reading level described themselves as being able to read or write English “well” or “very well.”209 In addition, low-literacy individuals are frequently ashamed of their reading difficulties and try to hide the problem, even from their families.221,222 Lack of recognition and denial of reading problems create a barrier to health care. Some low-literacy patients have acknowledged avoiding medical care because they are ashamed of their reading difficulties.221,222 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).221,222

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.223 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.224–227

Addressing health literacy is now recognized as critical to delivering person-centered health care.228 It is an important component of providing quality health care to diverse populations, and it will be incorporated into the National Standards for Culturally and Linguistically Appropriate Services.229 For example, translating highly technical medical and legal language into easily understood language is challenging, whether into English or another language. Health literacy practices are also included in the National Quality Forum 2010 updated set of safe practices.76 A recent discussion paper summarized 10 attributes that exemplify a “health literate health care organization.”228 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 PROMs 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.10 In some research and practice-based initiatives, evaluating cross-cultural differences in PROMs is of interest. In all these applications, researchers must use unbiased questionnaires that can detect important differences among patients.206,230,231

Possible cultural differences in interpreting questions and in response styles may limit data pooling or may constrain comparisons across members of different cultural groups.232–234 Similarly, poor quality translations can produce
noncomparable language versions of PROMs. For a questionnaire to be suitable for use as an unbiased measure of a PRO, items in the questionnaire must perform similarly across different groups (i.e., they must be cross-culturally or cross-linguistically equivalent). Without assurances that the PROM 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, PROMs 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 measurement. The prevalence of disability, defined as specific functional or sensory limitations, is estimated at 47.5 million Americans, or 22 percent of the US population. People with a 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 simulation initiatives.

Common disabilities that can affect PROM 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, to address many of these barriers, those administering such measures have 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.

Universal design refers to designing products and environments in such a way that all people can use them, to the greatest extent possible, without
adaptation or specialization. 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 people pushing children in strollers, among others. An exhaustive examination of how to apply the principles of universal design to PROM 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 in Section 508 of the Rehabilitation Act Amendments of 1998 can maximize flexibility. 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.

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.

For instance, for people with impaired vision, one might consider using in-person or telephone interviews (advantages and disadvantages discussed in an earlier section), an interactive voice response 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 (text telephones) 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 nonmotoric (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.

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 PROMs poses another barrier to the integrity of valid PRO assessment. Often, detecting true change is associated with the phenomenon of *response shift*. This 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).”256,p.1532 A change in perspective over time may result in patients’ attending to PROMs in a systematically different way from one time point to another.257

Response shift serves as a barrier to PRO assessment for several important reasons. For example, it threatens longitudinal PRO assessment validity, reliability, and responsiveness.257–260 Response shift can complicate the interpretation of PROM scores; a change in a PROM may occur because of response shift, an effect of treatment, or both.261

Monitoring for response shift can aid PROM users in interpreting longitudinal PRO data.259 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 shift257 and it is confounded with recall bias.260 Structural equation modeling has also been proposed as a way to identify response shift, but it is sensitive only if most of the sample is likely to make response shifts.262 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.263 Although growth modeling enables users to detect both the timing and shape of response shift,259 it cannot differentiate between random error and response shift.260
Implications of the Different Methods and Modes for Response Rate, Reliability, and Validity

Implementing Data Collection Methods

Users of PROMs must make a variety of decisions about the data collection method and the implications of those decisions on costs and errors in surveys. Two basic issues underlie these decisions: What is the most appropriate method to choose for a particular question, and What is the impact of a particular method on survey errors and costs?

Methods differ along a variety of dimensions. These include, although are not limited to, the degree of interviewer involvement and the level of interaction with the respondent. Channels of communication (sight, sound, touch) 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 Different Method or Mode Than the One Originally Validated

Considering the implications of using a different method or mode than the one on which the PROM was originally validated is also important. Many existing PROMs were initially validated in paper-and-pencil form. However, potential differences exist between paper-and-pencil and electronic-based PROM administration, 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 platforms).

As noted earlier, a growing body of research suggests measurement equivalence between paper- and computer-administered PROMs. However, the effect of a particular data collection method on a particular source of error may depend on the specific combination of methods used. Thus, as new methods are developed, studies comparing them with the methods they may replace must be done.

In framing expectations about the likely effect of a particular approach, developers need to invoke theories about that approach. Theory is informed by past mode-effects literature and by an understanding of the features or elements of a particular design. Similarly, mode choices involve trade-offs and compromises. Therefore, the choice of a particular approach must be made within the context of the particular objectives of the survey and the resources available.
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, optimization of response rates. 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. As in the case of using a different method or mode than the one in which the PROM was originally validated, instruments and procedures should be designed with an eye to ensuring equivalence across both methods and modes.

Accounting for the Impact of Nonresponders

Difficulties with data collection and questionnaire completion are major barriers to the successful implementation of PRO assessment. The principal problem is that missing data can introduce bias in analyses, findings, and conclusions or recommendations. The choice of mode and method of questionnaire administration can affect nonresponse rates and nonresponse bias. 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 nonresponse (one or more missing items within a questionnaire) or unit nonresponse (the whole questionnaire is missing for a patient). Evaluating the amount of, reasons for, and patterns of missing data is important. Some common strategies to evaluate nonresponse bias include:

- conducting an abbreviated follow-up survey with initial nonrespondents
- comparing characteristics of respondents and nonrespondents
- comparing respondent data with comparable information from other sources
- comparing on-time vs. late respondents

When dealing with missing data, analysts can use various statistical methods of adjustment. For item nonresponse in multi-item scales, several useful techniques tend to yield unbiased estimates of scores: simple mean imputation, regression imputation, and IRT models. For both item and unit
Patient-Reported Outcomes in Performance Measurement

nonresponse, 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).\textsuperscript{266,267} For unit nonresponse, users can implement a range of statistical techniques, depending on the reason for missing data.\textsuperscript{274-278}

Selection of Patient-Level PROMs

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. 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 disease and treatment affects patients’ health-related quality of life (HRQL). For such HRQL endpoints, in most cases, the patients are 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, health care delivery, and policy settings.

Access to psychometrically sound and decision-relevant PROMs will allow clinicians, investigators, administrators, and others to collect empirical evidence on the differential benefits and harms of a health-related intervention.\textsuperscript{279-282} Those obtaining such information can then disseminate findings to patients, clinicians and health care professionals, payers or insurers, and policy makers. Doing so may provide a richer perspective on the net impact of interventions on patients’ lives using endpoints that are meaningful to the patients.\textsuperscript{283}

Increasingly, longitudinal observational and experimental studies have included PROMs. To optimize decision making in clinical care, users must assess these PROMs in a standardized way, using questionnaires that demonstrate specific measurement properties.\textsuperscript{279,282,284-287} Our group recently identified minimum standards for the design or selection of a PROM for use in PCOR activities.\textsuperscript{288} Central to this work was understanding which attributes would make a PROM 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 PROMs and PCOR efforts to seek consensus on the minimum standards.\textsuperscript{288}

**Attributes of PROMs**

Many documents summarize attributes of a good HRQL measure. They include (an illustrative list) guidance documents from the FDA;\textsuperscript{289–292} the 2002 Medical Outcomes Trust guidelines on attributes of a good HRQL measure;\textsuperscript{293} the extensive, international expert-driven recommendations from COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments);\textsuperscript{285,294–298} the EORTC (European Organization for Research and Treatment of Cancer) guidelines for developing questionnaires;\textsuperscript{299} the Functional Assessment of Chronic Illness Therapy (FACIT) approach;\textsuperscript{30} the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force recommendation documents;\textsuperscript{149,241,300,301} and several others.\textsuperscript{245,284,302–304} Since 2010, ISOQOL (the International Society of Quality of Life) has completed two important guidance documents on use of PROMs in comparative effectiveness research and on integrating PROMs in health care delivery settings.\textsuperscript{284,305} Finally, the NIH PROMIS network released a standards document in 2012 that is useful for informing the minimal and optimal standards for designing PROMs.\textsuperscript{306}

Table 4 presents long-established criteria to consider in selecting PROMs for research, quality improvement activities, and now performance measurement. It specifies issues that PROM users need to consider when contemplating incorporating PROMs into performance measures and offers some best practices for evaluating PROMs in this context.

The eight primary criteria are the following: (1) conceptual or measurement model; (2) reliability and its subparts (e.g., internal consistency reliability); (3) validity and its subparts (e.g., content validity); (4) how scores are interpreted; (5) burden placed on respondents; (6) alternative modes and methods of administration; (7) cultural and language adaptations; and (8) use of electronic health records (EHRs). The table does not specify key issues and best practices for reliability or validity; that information is given only for the subcriteria. We illustrate these points with selected information pertaining to the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC).\textsuperscript{307}
Table 4. Primary criteria for evaluating and selecting patient-reported outcome measures (PROMs) for use in performance measurement

<table>
<thead>
<tr>
<th>Criteria and Subcriteria for Evaluating PROMs</th>
<th>Specific Issues to Address for Performance Measures and Best Practices in Assessing Candidate PROMs</th>
<th>Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)\textsuperscript{307} for Use in Hip Arthroplasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conceptual and Measurement Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Documentation should define and describe the concept(s) included and the intended population(s) for use.</td>
<td>• Target PRO concept should be a high priority for the health care system and patients. • Patient engagement should define what an important concept to patients is. • Target PRO concept must be actionable in response to the health care intervention.</td>
<td>• Patient input was used to evaluate the dimensionality and the importance of concepts to be measured.\textsuperscript{308} • Evidence suggests that some items measure the theoretically different concepts of physical function and pain load together on the same factor.\textsuperscript{309}</td>
</tr>
<tr>
<td>• Documentation should explain how the concept(s) are organized into a measurement framework, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationships among concepts.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Reliability</td>
<td>Adequate levels of reliability are prerequisites for determining the potential use of any PROM.</td>
<td>See 2a and 2b for WOMAC examples.</td>
</tr>
<tr>
<td>2a. Internal consistency (multi-item scales)</td>
<td>Classical test theory (CTT) typically relies on the following values: • Reliability estimate ≥ 0.70 for group-level scores • Reliability estimate ≥ 0.90 for individual-level scores Item response theory (IRT) typically uses the following: • Item information curves that demonstrate precision\textsuperscript{189} • A formula that can be applied to estimate CTT reliability.</td>
<td>Cronbach’s alphas for the three subscales (pain, stiffness, and physical function) range from 0.86 to 0.98.\textsuperscript{310–312}</td>
</tr>
</tbody>
</table>

(continued)
### Table 4. Primary criteria for evaluating and selecting patient-reported outcome measures (PROMs) for use in performance measurement (continued)

<table>
<thead>
<tr>
<th>Criteria and Subcriteria for Evaluating PROMs</th>
<th>Specific Issues to Address for Performance Measures and Best Practices in Assessing Candidate PROMs</th>
<th>Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) for Use in Hip Arthroplasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Reliability (continued)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2b. Reproducibility (stability over time)</td>
<td>Documentation should be provided about the specific reproducibility estimate used and the justification for the length of time between assessments.</td>
<td>For evaluating trends or changes over time, an adequate level of reproducibility is a prerequisite for determining the potential use of any PROM. Test-retest reliability has been adequate for the pain and physical function subscales, but less adequate for the stiffness subscale.</td>
</tr>
<tr>
<td>3. Validity</td>
<td>Documentation should explain the degree to which the instrument reflects what it is supposed to measure.</td>
<td>A limited number of PROMs have been validated for performance measurement. PROMs should include questions that are patient-centered. See 3a, 3b, and 3c for WOMAC examples.</td>
</tr>
<tr>
<td>3a. Content Validity</td>
<td>Documentation should explain the extent to which a measure samples a representative range of the content that it is supposed to cover, whether for the populations, settings, or other elements of the measurement task.</td>
<td>A PROM should have evidence supporting its content validity, including evidence that patients or experts (or both) consider the content of the PROM relevant and comprehensive for the concept, population, and aim of the measurement application. Meeting this criterion may entail: Development involved expert clinician input, survey input from patients, and a review of existing measures.</td>
</tr>
</tbody>
</table>

(continued)
Table 4. Primary criteria for evaluating and selecting patient-reported outcome measures (PROMs) for use in performance measurement (continued)

<table>
<thead>
<tr>
<th>Criteria and Subcriteria for Evaluating PROMs</th>
<th>Specific Issues to Address for Performance Measures and Best Practices in Assessing Candidate PROMs</th>
<th>Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)³⁰⁷ for Use in Hip Arthroplasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>3b. Construct and Criterion-Related Validity</td>
<td>A PROM should have evidence that</td>
<td>Patient ratings of satisfaction with arthroplasty were correlated with WOMAC scores in the expected direction.³⁰,³¹³, ³¹⁴</td>
</tr>
<tr>
<td>Documentation should explain how the PROM meets standard requirements for these two types of validity, giving appropriate evidence (empirical findings).</td>
<td>• supports predefined hypotheses about the expected associations among measures that are similar to or dissimilar from the measured PRO.</td>
<td>• Scores differentiated between patients with better versus worse outcomes after knee arthroplasty³⁰ and between patients with less versus more severe osteoarthritis.³¹⁵</td>
</tr>
<tr>
<td></td>
<td>• supports predefined hypotheses of the expected differences in scores between or among “known” groups.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• shows the extent to which scores of the instrument are related to a criterion measure.</td>
<td></td>
</tr>
<tr>
<td>3c. Responsiveness</td>
<td>For longitudinal initiatives or applications, documentation should explain how the PROM can detect change over time and change in response to an intervention (i.e., empirical findings of changes in scores consistent with predefined hypotheses regarding changes in the target population).</td>
<td>Responsiveness and ability to detect change in response to clinical intervention are both adequate.³¹⁶</td>
</tr>
<tr>
<td></td>
<td>• If a PROM has cross-sectional data that provide 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 PROM to detect changes in the construct being measured over time), users need to consider carefully whether such a measure is likely to provide valid data over time in a longitudinal study, especially if no other PROM is available.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Emphasizing responsiveness is important because of the expectation that care will have consequences. If action is to be taken, then demonstrating responsiveness is important.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• PROMs must be sensitive to detect change in response to the specific health care intervention.</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
### Table 4. Primary criteria for evaluating and selecting patient-reported outcome measures (PROMs) for use in performance measurement (continued)

<table>
<thead>
<tr>
<th>Criteria and Subcriteria for Evaluating PROMs</th>
<th>Specific Issues to Address for Performance Measures and Best Practices in Assessing Candidate PROMs</th>
<th>Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) for Use in Hip Arthroplasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Interpretability of Scores</td>
<td>Documentation should support and assist users in interpreting scores from the PRO measure, including:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• If different PROMs are used, establishing a link or crosswalk between them is important.</td>
<td>• Population-based, age- and sex- (or gender-) normative values are available.</td>
</tr>
<tr>
<td></td>
<td>• Because the criteria for assessing clinically important change in individuals do not directly translate to evaluating clinically important group differences, a useful strategy is to calculate the proportion of patients who experience a clinically significant change.</td>
<td>• Minimal clinically important improvement values are available.</td>
</tr>
<tr>
<td></td>
<td>• Guidance on the minimally important difference in scores between groups or over time (or both) that can be considered meaningful from patient and clinical perspectives.</td>
<td>• Instrument can be translated into a utility score for use in economic and accountability evaluations.</td>
</tr>
<tr>
<td></td>
<td>• If different PROMs are used, establishing a link or crosswalk between them is important.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• If different PROMs are used, establishing a link or crosswalk between them is important.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Representative mean(s) and standard deviation(s) in the reference population</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Guidance on the minimally important difference in scores between groups or over time (or both) that can be considered meaningful from patient and clinical perspectives.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• In a busy clinic setting, PRO assessment should be as brief as possible, and reporting should be done in real time.</td>
<td>• Short form is available.</td>
</tr>
<tr>
<td></td>
<td>• Patient engagement should inform what constitutes “burden.”</td>
<td>• Average time to complete mobile phone WOMAC is 4.8 minutes.</td>
</tr>
<tr>
<td>5. Burden</td>
<td>In a busy clinic setting, PRO assessment should be as brief as possible, and reporting should be done in real time.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Population-based, age- and sex- (or gender-) normative values are available.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Minimal clinically important improvement values are available.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Instrument can be translated into a utility score for use in economic and accountability evaluations.</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Table 4. Primary criteria for evaluating and selecting patient-reported outcome measures (PROMs) for use in performance measurement (continued)

<table>
<thead>
<tr>
<th>Criteria and Subcriteria for Evaluating PROMs</th>
<th>Specific Issues to Address for Performance Measures and Best Practices in Assessing Candidate PROMs</th>
<th>Example: The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)307 for Use in Hip Arthroplasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Cultural and Language Adaptations</td>
<td>Documentation should describe methods to evaluate cultural and linguistic equivalence.</td>
<td>The mode, method, and question wording must yield equivalent estimates of PRO measures.</td>
</tr>
<tr>
<td></td>
<td>Instrument is available in more than 65 languages.</td>
<td></td>
</tr>
<tr>
<td>8. Electronic Health Records</td>
<td>Documentation should describe key considerations for incorporation into electronic health records.</td>
<td>Critical features include:</td>
</tr>
<tr>
<td></td>
<td>• Interoperability</td>
<td>Electronic data capture may allow for integration within electronic health records.22</td>
</tr>
<tr>
<td></td>
<td>• Automated, real-time measurement and reporting</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Sophisticated analytic capacities.</td>
<td></td>
</tr>
</tbody>
</table>

Note: This table is adapted from recommendations in a report from the Scientific Advisory Committee of the Medical Outcomes Trust293 and a report submitted to the Methodology Committee of the Patient Centered Outcomes Research Institute.288 We adapted the key points from these sources to enhance relevance to PRO selection for performance measurement.

Important Differences in PROM Attributes

Selecting PROMs 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 PROMs for research purposes rather than these other nonresearch purposes. Generally speaking, the factors to consider when selecting PROMs for performance measurement and quality improvement activities are more similar than different. Thus, we focus here more on the differences that users will need to take into account.

Instrument Length

One key difference involves the length of the PROM. Longer questionnaires may be better tolerated in the context of research than in clinical practice settings; thus, to facilitate widespread adoption, PROMs for performance measurement should be short surveys. Addressing the need for shorter PROMs may, however, compromise other important measurement characteristics, such as reliability (i.e., precision and reproducibility).
Implications of PRO Data for Action

Another key difference in factors to consider when selecting PROMs for clinical practice quality improvement, or performance measurement and accountability efforts, is the implications or consequences of the PRO data. Specifically, using PROMs 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 PROMs are higher in the performance measurement context than in research applications.

The problem lies, in part, in the constraints to the quality of the measurement level arising from factors unique to performance measurement. These can include instrument length or representativeness of the patient or consumer populations surveyed. These considerations highlight the importance of emphasizing responsiveness and sensitivity to change when considering PROMs for use in the ways envisioned for NQF-endorsed measures.

History of Successful Use of PROMs

In selecting a PROM for these various purposes, a logical first step involves reviewing what measures have already been used successfully. Using PROMs for these programs remains an understudied area, but several examples of PROMs used as indexes of performance measurement provide an initial foundation upon which the field can expand.

The Veterans Health Study assessed PROs within the Veterans Health Administration (VA) system. In response to the VA’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 health care system levels. The Centers for Medicare & Medicaid Services (CMS) and its Medicare Advantage Program have applied these methods for similar purposes, and CMS has also designated the VR-12 as the principal outcome measure of the Medicare Health Outcomes Survey (HOS).

Research examining the VR-36 and SF-36 in such uses does inform the selection of PROs for performance measurement. Nevertheless, limitations remain to 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 the “static” nature of these measures, meaning that for analysts to be able to obtain an individual’s score, all items must be administered—even those items that 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 and “dynamic” instruments 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 PROMs 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.187 PROMIS measures form a hybrid between static generic PROMs and more flexible adaptive measures. They comprise items that are specific to the overall content of the measure but that are also applicable across the diverse spectrum of health status.

Although a growing body of literature provides preliminary evidence supporting the psychometric quality of the PROMIS measures, future work needs to explore applying PROMIS measures as tools for assessing the performance of health care organizations. Nevertheless, the PROMIS system provides a robust model by which the use of PROMs as performance measures can be expanded and elaborated upon, owing to its rigorous methodological characteristics.

**Documentation of Particular Attributes of PROMs**

Documentation, in peer-reviewed literature or on publicly accessible websites (or both), of the evidence of a PROM to reflect all of these measurement properties will improve acceptance of the PROM for use as a performance measure. To the extent that the evidence came from populations similar to the studies’ target populations, the more confidence clinicians, analysts, administrators, and policy makers can have in the PROM 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 people selected to participate in the various pilot tests or
studies that are part of the evaluation of the measurement properties for the PROM. 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 identifying the domains most important to measure and in selecting specific PROMs for use in performance measurement.

Participants’ literacy is another important consideration for use of PROMs. Data collected from PROMs are valid only 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. Developers of PROMs must ensure that the questions and response options are clear and easy to understand. Pretesting of the instrument (e.g., cognitive testing) should include individuals with low literacy to evaluate the questions.

Response burden must be considered when selecting a PROM. The instrument must not be overly burdensome for patients, as they are often sick and cannot be expected to tolerate completing lengthy questionnaires.

Finally, researchers must carefully consider the strength of evidence for the measurement properties. No threshold exists to indicate that an instrument is (or is not) valid for any or all populations or applications. In addition, no single study can confirm 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 PROM. More established PROMs will have the benefit of having accrued more evidence than more recent entries; however, more recent entries tend to have improved measurement properties that warrant attention.

**PROM Characteristics for Consideration**

**Generic Versus Condition-Specific Measures**

One factor to consider when selecting a patient-level PROM is whether to use a generic instrument or a condition-specific instrument. Several considerations can inform this choice. 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, or people with multiple comorbidities, a generic measure is the preferred choice. Conversely, if the goal is to examine a specific subset of patients with a particular diagnosis
or receiving a common treatment, then a condition-specific measure may be more appropriate, but this is ideally evaluated in context.

In addition, outcomes of interest may guide the selection process. Generic measures may capture a different category of outcomes when compared with a condition-specific PROM. For example, a generic measure may assess domains of general function, well-being, or quality of life, whereas a condition-specific PRO may measure symptoms expected to be directly addressed by a condition-specific intervention. The more focused the interest in a specific symptom or set of symptoms that are unique to the condition, the more likely a condition-specific instrument will be preferred.332

Generic PROMs have some important advantages. They allow for comparability across patients and populations,331 although they are more suitable for comparison across groups than for individual use.333 Global PROMs also allow assessments in terms of normative data that can be used to interpret scores.331 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.331

Generic PROMs also have some disadvantages. They may tend to be less responsive than condition-specific measures to focal changes that are better detected with a condition-specific measure. For that reason, they may underestimate health changes in specific patient populations.334 Additionally, they may fail to capture important condition-specific concerns.334

Condition-specific PROMs are an alternative to generic PROMs. One advantage of condition-specific PROMs is the possibility for improved relevance and responsiveness.331 They also enable differentiation of groups at the level of specific symptoms or patient concerns.331 However, the condition-specific focus introduces the notable difficulty of making comparisons across patient populations with different diseases or health conditions.331

Given their respective benefits and limitations, we recommend that a combination of generic and condition-specific measures is likely to be the best choice for the performance measurement purposes that those assessing or reporting on quality of care in this country, such as the NQF, have most in mind. Generic and condition-specific PROMs may measure different aspects of HRQL when administered in combination,335 resulting in more comprehensive assessment. Consequently, hybrid measurement systems have emerged to facilitate combining them. For example, the FACIT system consists
Method and Mode of Administration, Data Collection, and Analysis

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, represents another example of a hybrid system of PROMs that combines both global and targeted approaches.

**Measurement Precision**

Another factor to consider when selecting a patient-level PROM 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. PROMs with greater measurement precision also demonstrate greater sensitivity to change. Given that most PROMs were originally developed as research tools, they may lack the level of precision necessary for assessing individuals on these types of outcomes. Although performance measures will aggregate to practice, provider, or organization levels, adequate measurement precision at the patient level is still needed.

Regarding measurement precision, measures based on IRT tend to have greater precision than measures based on classical test theory. 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. 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 PROMs derived from IRT techniques is recommended to achieve the greatest measurement precision.

**Sensitivity to Change, or Responsiveness**

Sensitivity to change (also referred to as responsiveness) is another important factor to consider when selecting a PROM because the ability to detect a small, but important, change is necessary when monitoring patients and implementing clinical interventions. Sensitivity to change is a type of validity characterized by within-subject changes over time following an intervention.

Responsiveness is conceptualized in many ways, which leads to different findings and interpretations. 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, PROMs must demonstrate
sensitivity to change both when individuals improve and when they deteriorate.342

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.343 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.344 Indeed, relying solely on statistical tests of responsiveness is not recommended, given that such findings may not accurately reflect what is meaningful to patients or clinicians.345

Several factors can limit responsiveness to change. First, multi-trait scales containing items that are not relevant to the population being assessed may fail to capture change over time.346 The responsiveness of a PROM may also be constrained by using scales that offer categorical or a limited range of response options.346 PROMs 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.346 The responsiveness of a PROM is also limited when it includes items that reflect stable characteristics that are unlikely to change. Scales that contain items with floor or ceiling effects are also problematic.346 A PROM sensitivity to change may depend upon the direction of the change. For example, Eurich and colleagues found that PROMs were more responsive to change when patients got better clinically than when they got worse.38

In addition to these factors, a growing body of research suggests that condition-specific PROMs can be more sensitive to change than generic PROMs.38,40,347–349 Responsiveness to change is likely influenced by the purpose for which the measure was originally developed.349 For example, measures developed to emphasize specific content areas would be expected to show greater post-treatment change in those content areas.342 The greater sensitivity to change in condition-specific PROMs may be attributed to the strong content validity inherent in condition-specific measures.38 As a result, using a combination of condition-specific and generic PROM may yield the most meaningful data.38,40
Minimally Important Differences

The difference between clinical versus statistical significance also merits consideration when selecting a PROM. Historically, research has relied upon tests of statistical significance to examine differences in 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 PROM scores.

Experts have proposed a variety of approaches to determining clinical significance. For example, clinically significant change has been defined as “changes in patient functioning that are meaningful for individuals who undergo psychosocial or medical interventions.” Similarly, meaningful change is defined (from the patient perspective) as “one that results in a meaningful reduction in symptoms or improvement in function . . . .”

Minimally important differences (MIDs) represent a specific approach to clinical significance. They are defined as “the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important.” Minimum clinically important differences (MCIDs) constitute an even more specific category of MID. MCIDs 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.”

Examining clinically significant differences poses several important implications. First, investigating clinically significant (versus statistically significant) differences in scores aids users in interpreting PROMs. Second, focusing on clinically significant differences also emphasizes the importance of the patient perspective, which may not be adequately captured when looking mainly at statistically significant differences. Third, the ability to look at clinically significant differences in 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. 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 PROM and some independent measure that is interpretable. Evaluators have several options for the type of anchor they...
might select when using an anchor-based method. For instance, clinical anchors that are correlated with the PROM at the $r \geq 0.30$ level may serve as appropriate anchors. Clinical trial experience can inform the selection of these clinical anchors, including the use of multiple clinical anchors.

Transition ratings represent another potential source of anchors when establishing MID s. Transition ratings are patients’ within-person ratings of change. However, because of concerns about validity, experts recommend that researchers or other users examine the correlation between pre- and post-test scores and the transition rating. Patients’ between-person differences can also be used as anchors when establishing MID s for PROMs. Additional sources for anchors when establishing MID s include HRQL-related functional measures used by clinicians and objective standards (e.g., hospital admissions, time away from work).

Although the anchor-based method offers promise for establishing MID s in PROMs, several limitations should be considered. First, the transition rating approach to anchor selection is subject to recall bias on the part of the patient. Second, global ratings may account for only some variance in scores. Third, the anchor-based method does not take into consideration the measurement precision of the instruments being used.

The distribution-based method represents the second method of establishing MID s in PROMs. The distribution-based method uses the statistical characteristics of the scores when establishing MID s. Specifically, the distribution-based approach evaluates change in scores in relation to the probability that the change occurred at random.

As in the case of the anchor-based method, several methods are available when applying a distribution-based approach to establishing MID s. First, the $t$-test statistic has been used to establish MID s when examining change over time. 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. Second, distribution-based methods may also be grounded in measurement precision and the standard error of the mean (SEM). Specifically, the 1 SEM criterion can be used as an alternative to MID when assessing the magnitude of PROM score changes. Sample variation, such as effect size and standardized response mean, constitutes another method for establishing MID s using the distribution-based method. When using this method, it is recommended that the effect size be specific to the population being studied. Evidence suggests that MID estimates using sample variation are approximately one-half of a standard deviation.
Finally, reliable change constitutes another method of using the
distribution-based approach to establish MIDs. Reliable change is based
on the standard error of measurement difference (SEMD); it indicates how
much the observed change in an imprecise measure exceeds fluctuations that
are random in nature. Although the distribution-based approach serves as a
possible alternative to the anchor-based methods, little consensus exists on the
benchmarks for establishing changes that are clinically significant.

Given limitations of the anchor- and distribution-based approaches, experts
recommend that users apply multiple methods and triangulation to determine
the MID. Moreover, the final selection of MID values should be based
on systematic review and an evaluation process such as the Delphi method.
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, in some cases, the desired outcome may be
scores over time, such as in the case of interventions designed to preserve and
prevent declines in functioning. Consequently, the specific application of the
PRO will inform the 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 PROMs, evaluators should not apply a
single MID to all situations. MIDs may vary by population and by context.
Consequently, those reporting such data should provide a range around the
MID, rather than just a single MID value. Finally, because the criteria for
assessing clinically important change in individuals do not directly translate
to evaluating clinically important group differences, a useful strategy is to
calculate the proportion of patients who experience a clinically significant
change.

**Essential Conditions to Integrate PROMs Into the
Electronic Health Record**

**General Considerations for Health Information Technology**

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

*E-health* refers to health-related Internet applications that deliver a range
of content, connectivity, and clinical care. This includes health information,
online formularies, prescription refills, appointment scheduling, test results,
advance care planning and health care proxy designation, and physician-patient communication.\textsuperscript{362} Patient-centered e-health (PCEH) is an emerging discipline that is defined as the combination of three themes:\textsuperscript{363}

- 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 providers.\textsuperscript{364} However, little evidence is available as to whether the services offered by providers are services that patients desire.\textsuperscript{12} One important consideration is that providers attend to patient acceptability factors.\textsuperscript{12,365}

Measuring PROMs will constitute an important aspect of future stages of “meaningful use” of electronic health records (EHRs).\textsuperscript{366,367} Access can be enhanced by allowing entry directly from commonly used devices such as smartphones. Enabling clinical decision support by providing structured data directly into EHRs will permit PROMs to be used for (1) tracking patient progress over time or (2) through individual question responses, driving 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 [Logical Observation Identifier Names and Codes]) enables EHRs to incorporate the instrument as an observation with a known set of responses using standard terminology (SNOMED-CT [Systematized Nomenclature of Medicine—Clinical Terms]) 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 HIT studies published between 2004 and 2007, PROMs were not mentioned at all.\textsuperscript{362}

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 the 2010s. 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.

In summary, important issues include (1) the patient perspective (patients want to be involved “as a participant and partner in the flow of information” relating to their own health care); (2) clinical buy-in; (3) compatibility with clinical flow; and (4) meaningful use.

Examples of PROMs in Electronic Health Record Applications

Health care centers are beginning to implement ways to use patient-reported information (the voice of the patient) to provide higher quality care. Three recent case studies (two in the United States and one in Sweden) are particularly informative, because they illustrate lessons learned about such initiatives.

The Dartmouth Spine Center collects health survey data from patients before each visit, either at home or in the clinic. Analysts summarize the data in a report and make it available for use by patients and clinicians to develop or modify care plans and to monitor results over time to guide treatment decisions. Longitudinal changes are incorporated into the report with each new assessment. At Group Health Cooperative in the State of Washington, an electronic health risk assessment has been integrated with the EHR. Patients can complete PROMs, make appointments, fill prescriptions, review health benefits, communicate with their providers, and get vetted health information. Customized reports are available to patients and providers. The Karolinska University Hospital in 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. Beginning in 2003, its web-based system replaced paper forms. The system uses real-time data provided by patients, clinicians, and diagnostic tests. Longitudinal summaries of PROMs and other health information are incorporated into graphical reports that 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 PROMs 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 PROMs into the flow of care, designing the systems with stakeholder engagement, merging 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 include use of PROMs in managing 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 HRQL. Integrating formal HRQL assessment into treatment decision making can improve patient-centered care for cancer patients with advanced disease. Computer-based assessment can reduce patient and administrative burden while enabling real-time scoring and presentation of HRQL data. Two pilot studies conducted with patients with advanced lung cancer reported that the computer technology was acceptable and feasible for patients and physicians. 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 at Northwestern University involves developing and implementing patient-reported symptom assessment in gynecologic oncology clinics. Before their clinic visits, outpatients complete instruments measuring fatigue, pain, physical function, depression, and anxiety through the EHR patient communication portal at home or in the clinic using an iPad. Results immediately populate the EHR. Severe symptoms trigger EHR notifications to providers. The EHR also provides automated triage for psychosocial and nutritional care when indicated.
Selection of PROMs That Meet Recommended Characteristics for Use in Performance Measures

Throughout this monograph, we have recommended several criteria that researchers and evaluators can use when assessing the appropriateness of a PROM for measuring quality of care and performance; Table 4 summarized critical points. Given that PROMs are not yet in widespread use in clinical practice, little is known about how best to aggregate these patient-level outcomes for measuring the quality of care or performance of the health care entity. Despite this limitation, accommodating the needs of patients with diverse linguistic, cultural, educational, and functional skills calls for evidence about the equivalence of multiple methods and modes of questionnaire administration. Additionally, scoring, analyzing, and reporting PRO response data all need to be user-friendly and understandable to clinicians for real-time use in clinical settings. Moreover, the timing of measurement must include administration before therapeutic interventions to allow for measuring responsiveness to change, doing risk adjustment, and screening patients for clinical intervention.

To illustrate the application of these recommended characteristics when evaluating the appropriateness of a PROM for these purposes, Table 4 included one illustration of these points related to determining the success of total hip arthroplasty. Total hip arthroplasty has emerged as an acceptable surgical treatment for individuals experiencing intractable pain and severe functional impairments for whom conservative treatment has yielded minimal improvement.373–376 The most common indication for total hip arthroplasty is joint deterioration secondary to osteoarthritis.377 Consequently, the aging of the population is likely to raise demand for both primary total hip arthroplasty and revision procedures.378–380

PROs have increasingly been included alongside more traditional indices of surgical outcome such as morbidity and mortality when evaluating the success of total hip arthroplasty. With the expanding focus on patient-reported outcomes, such as functioning and quality of life, numerous, diverse PROMs have been developed and applied in measuring total hip arthroplasty outcomes.377 Thus, this intervention provides a relevant context in which to review the use of recommended characteristics in the selection of PROMs, specifically with the characteristics of the WOMAC, a PROM developed to examine pain, stiffness, and physical function in individuals with osteoarthritis.307
Conclusions

PRO measures have reached a level of sophistication that enables wider use in assessing performance in clinical settings. Attention to the many methodological considerations discussed in this monograph will help users to produce meaningful, actionable results. Judicious use of a mixture of generic and condition-specific assessment instruments, acceptance of modern measurement methods such as IRT, and application of technology to enable standardized, equitable assessment across a range of patients are essential in this process. Implementing contemporary options, such as those offered by the PROMIS instruments, can effectively shorten assessment time without compromising accuracy. These attributes facilitate meeting the demands of clinical application of PROs for performance measurement.
References


199. Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care. 2000;38(9 Suppl):II60-II5.


306. PROMIS Validity Standards Committee on behalf of the PROMIS Network of investigators. The PROMIS instrument development and psychometric evaluation scientific standards. 2012.


Acknowledgments

The authors thank Kathleen Swantek, MLIS, for assistance with reference management. We also acknowledge our appreciation for helpful comments on the content of an earlier draft of this work from Karen Adams, PhD, MT; Ethan Basch, MD, MSc; Victor Chang, MD; Stephan Fihn, MD, MPH; Floyd Jackson Fowler, PhD; Lewis Kazis, ScD; Jennifer Moore; Eugene Nelson, DSc, MPH; Kenneth Ottenbacher, PhD, OTR; Karen Pace, PhD, MSN; and Mary Tinetti, MD. We also acknowledge the expert assistance of Joanne Studders, BA, and Loraine Monroe for diligent editing and document preparation, respectively.
About the Authors

Zeeshan Butt, PhD, is an Associate Professor of Medical Social Sciences, Surgery, and Psychiatry and Behavioral Sciences at Northwestern University Feinberg School of Medicine. He also serves as the Associate Director of the Center for Patient-Centered Outcomes there.

David Cella, PhD, is a Professor and Chair of the Department of Medical Social Sciences at Northwestern University Feinberg School of Medicine. He is the principal investigator of the National Institutes of Health (NIH) Initiative “HealthMeasures: A Person-Centered Assessment System,” which curates and distributes PROMIS (the Patient Reported Outcomes Measurement Information System) and three other health measurement systems. Dr. Cella is also a founding member and officer of the PROMIS Health Organization, a charitable organization that licenses and promotes the use of PROMIS worldwide.

Elizabeth A. Hahn, MA, is an Associate Professor in the Departments of Medical Social Sciences and Preventive Medicine, and the Institute for Public Health and Medicine, at Northwestern University Feinberg School of Medicine. Her research primarily involves patient-reported outcomes in chronic illnesses, with a focus on underserved populations and health disparities.

Sally E. Jensen, PhD, is a Research Assistant Professor in the Departments of Medical Social Sciences and Surgery at Northwestern University Feinberg School of Medicine. Her research focuses on patient-reported outcomes across various medical and surgical patient populations.

Kathleen N. Lohr, PhD, is a Distinguished Fellow at RTI International. Her research has involved comparative effectiveness and evidence-based practice, measurement of health status and quality of life, quality of care assessment, clinical practice guidelines, and related health policy issues.

Cindy J. Nowinski, MD, PhD, is a Research Associate Professor in the Departments of Medical Social Sciences and Neurology at Northwestern University Feinberg School of Medicine. Her research focuses on the development of patient-reported and objective health outcome measures for use in clinical and epidemiological research and clinical care.

Nan Rothrock, PhD, is a Research Associate Professor in the Department of Medical Social Sciences at Northwestern University Feinberg School of Medicine. Her research focuses on development and application of patient-reported outcome measures in research and practice.
Praise for *Patient-Reported Outcomes in Performance Measurement*

“Here it is, all in one place. A compact book on patient-reported outcomes beginning with the idea and ending with implementation for performance measurement and integration in electronic health records. A great resource from a team of experts.”

—Eugene C. Nelson, DSc, MPH, Professor, The Dartmouth Institute for Health Policy and Clinical Practice

“NQF supported this seminal work to help amplify the essential patient voice in quality measurement. This work provides the right balance of science and practical advice needed to move patient-reported outcomes into the mainstream of performance measurement.”

—Helen Burstin, MD, MPH, Chief Scientific Officer, National Quality Forum

“Measurement of patient-reported outcomes is the next frontier in health care delivery quality assessment. Although interest is rapidly growing to integrate the patient’s voice into performance evaluation, guidance is limited on how actually to go about doing this. This monograph provides invaluable practical and methodological insights for those aiming to integrate patient-reported outcomes into a quality program, ranging from selection of outcomes, to measures, analysis, and reporting.”

—Ethan Basch, MD, MSc, Director, Cancer Outcomes Research Program, University of North Carolina at Chapel Hill