HHS Public Access

Author manuscript

J Clin Epidemiol. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

J Clin Epidemiol. 2015 September ; 68(9): 1085–1092. doi:10.1016/j.jclinepi.2015.03.023.

Pragmatic Characteristics of Patient-Reported Outcome Measures are Important for Use in Clinical Practice

Kurt Kroenke[a,b,c,*], **Patrick O. Monahan**[d], and **Jacob Kean**[a,c,e]

[a] VA HSR&D Center for Health Information and Communication, Roudebush VA Medical Center, Indianapolis, IN

[b] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN

[c] Regenstrief Institute, Inc., Indianapolis, IN

[d] Department of Biostatistics, Indiana University School of Medicine and School of Public Health, Indianapolis, IN

[e] Department of Physical Medicine and Rehabilitation, Indiana University School of Medicine, Indianapolis, IN

Abstract

Objective—Measures for assessing patient-reported outcomes (PROs) that may have initially been developed for research are increasingly being recommended for use in clinical practice as well. While psychometric rigor is essential, this paper focuses on pragmatic characteristics of PROs that may enhance uptake into clinical practice.

Methods—Three sources were drawn upon in identifying pragmatic criteria for PROs: 1) selected literature review including recommendations by other expert groups; 2) key features of several model public domain PROs; 3) the author' experience in developing practical PROs.

Results—Eight characteristics of a practical PRO include: 1) actionability (i.e., scores guide diagnostic or therapeutic actions/decision-making); 2) appropriateness for the relevant clinical setting; 3) universality (i.e., for screening, severity assessment, and monitoring across multiple conditions); 4) self-administration; 5) item features (number of items and bundling issues); 6) response options (option number and dimensions, uniform vs. varying options, timeframe, intervals between options); 7) scoring (simplicity, interpretability); and 8) accessibility (nonproprietary, downloadable, available in different languages and for vulnerable groups, incorporated into electronic health records)

Conclusion—Balancing psychometric and pragmatic factors in the development of PROs is important for accelerating the incorporation of PROs into clinical practice.

[*]Corresponding author: Kurt Kroenke, MD, Regenstrief Institute, 5th Floor, 1050 Wishard Blvd, Indianapolis, IN 46202. Ph 317-630-7447 FAX 317-630-8776. kkroenke @regenstrief.org.

None of the authors have any conflicts of interest to disclose.

Keywords

patient-reported outcomes; psychometrics; utility; measures; scales

Measurement is a vital aspect of patient care, necessary for diagnosis, grading of disease severity, estimating prognosis, and monitoring and adjusting treatment. However, not all relevant outcomes can be assessed with a device, a laboratory test, a physical finding, or some other data gathered independent of the patient's perceptions and voice. Symptoms, health-related quality of life, and certain other domains rely exclusively or predominantly on patient-articulated feelings and experiences and therefore depend upon reliable and valid patient-reported outcome (PRO) measures. Indeed, the NIH has recognized the importance of PROs by investing heavily in the development of the Patient-Reported Outcome Measurement Information Systems (PROMIS) scales(1) freely available at www.promis.org.

In this paper, we propose several factors to consider when developing a practical PRO measure. By practical we mean those features that will enhance a measure's adoption and use in clinical practice. William James, a key founder of the pragmatic school of American philosophy, defined truth as that “which works” or has “cash value”.(2) The “cash value” of a PRO is its relevance to patient care.

The practical characteristics outlined in Table 1 do *not* include the classical psychometric requirements of a scale, such as reliability or validity, nor do they speak to the many basic and advanced procedures for scale development (e.g., item selection, cognitive testing, differential item functioning, item response theory). Psychometric standards are a given and well-described in consensus reports on PROs.(3,4) Indeed, most PROs gravitate from research into practice and practical considerations should not override the necessity for psychometric rigor in scale development. Table 2 compares our pragmatic recommendations with those of several other groups(5-7), although the latter groups may sometimes use alternative terms or raise different issues related to the 8 characteristics as well as suggest other practical considerations.

Scale development is often a low priority for sponsors that support biomedical research, thereby constraining the funding available for evaluating every psychometric nuance of a PRO. This is especially true when a measure is developed and “second-generation” questions arise, such as: 1) differences between modes of administration (e.g., self-report vs. interview; patient vs. proxy; in-person vs. telephone); 2) standards for translating into different languages; 3) abbreviating or modifying versions of the original measure. Therefore, we advocate a balance between psychometric and pragmatic values in all stages of PRO development and validation. The OMERACT guidelines exemplify a similar balance even for outcome measures used in clinical trials by not only recommending truth and discrimination as psychometric criteria but also feasibility (e.g., can the measure be applied easily, given constraints of time, money, and interpretability?) as a pragmatic criterion.(8)

1. Actionability

The utility of a PRO in clinical practice is enhanced when providers know how to translate scores into concrete actions, such as further diagnostic evaluation or testing, treatment initiation or adjustment, or subspecialty referrals.(5,9) Simply providing more data to busy practitioners who already have enormous competing demands for their time in a clinical encounter often limited to 15 minutes or less can be more frustrating than empowering. (10,11) On the other hand, data that efficiently informs specific actions will be embraced. For example, a high depression score prompting an increase in the antidepressant dose can be as useful as an elevated serum cholesterol that leads to modifying lipid-lowering therapy. A useful preference-based question asks patients if they desire treatment for their symptoms. (12,13) This provides a patient-centered criterion for interpreting PRO scores in the individual person, since different patients may desire (or alternatively refrain from) treatment at different symptom thresholds.

What factors might make a PRO <u>not</u> actionable in a particular clinical setting?

a) The target of the PRO may be *outside the purview* of a particular clinician who in turn lacks referral options. For example, social functioning is a domain many physicians neither have the skills nor resources to address. Thus, unless a social work referral or community resource are readily available, knowledge of impaired social functioning in the absence of explicit actions to efficiently address these impairments can be demoralizing for the clinician and offer false hope to the patient.

b) The target may be within the purview of the clinician but *resource-contingent*, such that in the absence of these resources, use of the PRO will not benefit patient outcomes. For example, multiple trials have shown that depression screening alone does not enhance outcomes(14) but depression screening combined with other systems enhancements does. (15) This has led the US Preventive Services Task Force to recommend use of a depression screening measure only if systems are in place to adequately optimize depression outcomes. (16)

c) The domain assessed by the PRO may be *excessively bundled*, in which case a particular score cannot inform a targeted action without efforts by the clinician to conduct a differential diagnosis of what may be leading to an elevated score and then determine what is and is not actionable. For example, a physical function or role function score may be abnormal due to numerous medical and nonmedical factors. However, there is no discrete “physical function” or “role function” pill, procedure, or other specific therapy. Still, such summary scores might be useful at a higher level (e.g., assessing quality of care or systems-based interventions provided to patient panels or populations).

2. Setting-appropriate

The clinical setting and the actionable goals of that setting may dictate the length, specificity, and granularity of a particular PRO.(7) For example, brief PROs capturing multiple domains may be more suitable for primary care, whereas longer PROs focused on one or several disease-specific domains may be preferable for specialty clinics. PROs can

also be used in non-office-based settings, such as hospitals or long-term care facilities. The rationale for shorter PROs in the general medical setting is not only because the clinician has to attend to more conditions but also because longer measures targeting multiple domains increase the respondent burden for the patient. For example, a 30-item PRO with 5 items for each of 6 domains that is used in primary care would have a comparable respondent burden as a 30-item cardiovascular disease-specific PRO completed by patients in a cardiology clinic. One caveat – longer disease-specific measures are only justified for a specialty setting if they lead to different or differential or distinctive diagnostic or treatment actions. For example, if a brief depression or pain measure performs as well as a longer one, it may suffice in all settings.

3. Universal

Measures can be used to screen for a condition, to establish a diagnosis, to assess severity, to monitor treatment response and to provide prognostic information about the future course of a disease.(5,17) After all, the same blood pressure cuff is used to screen for, diagnose, and assess the severity of hypertension as well as monitor the adequacy of blood pressure treatment. One example of a multi-purpose PRO is the PHQ-9 which can be used as to assess depression severity, establish probable diagnoses, and monitor treatment. (18) It is also valid across a range of ages (adolescent to geriatric populations)(19,20)as well as racial/ethnic, educational, and sociocultural factors.(18,21). Likewise, a “cross-cutting” measure that can be used in multiple conditions and diseases has certain advantages. Some symptoms (pain, fatigue, depression, sleep disturbances) occur in many medical and mental disorders, and patients frequently have a number of comorbid conditions. Similarly, functional status and health-related quality of life domains often transcend specific diseases. The degree to which a PRO is broadly applicable across multiple conditions increases its utility and obviates the need for multiple measures of the same domain to cover different diseases. Indeed, this has been a key principle in developing PROMIS measures.(22)

4. Self-administered

As opposed to measures of bodily fluids (e.g., laboratory tests of blood, urine, sputum) or organs (physical examination, radiographic imaging, pathological specimens), assessment of symptoms relies largely or exclusively on patient report. In this regard, self-administered measures can minimize the amount of time devoted to mere data collection allowing the clinician more time to discuss the meaning and treatment implications of elevated PRO scores. Second, self-administered measures may lead to greater detection of problems in sensitive areas like substance use or sexual function.(23) Third, self-administered measures are not influenced by clinician or interviewer biases that can occur when an external observer superimposes his or her interpretation of symptoms that only the patient can accurately grade. That being said, it is also desirable if a measure is reliable and valid across various modes of administration (e.g., self- vs. interviewer administered; in-person vs. telephone vs. web-based) or, in some cases, even different perspectives (e.g., a clinical observer or a proxy when assessing pain or depression in children or in adults with cognitive impairment).(20,24,25)

5. Items

Brevity is generally desirable; we arbitrarily define a brief measure as “single digits” (less than 10 items) and an ultra-brief measure as 1 to 4 items. Another metric might be the length of time it takes an individual to complete a PRO or a set of PROs – anything requiring more than 1 to 5 minutes may be perceived as excessive in busy outpatient settings . As noted, the minimal length of a measure may be dictated by the purpose (e.g., screening vs. monitoring), although there is some evidence that even ultra-brief PROs may be perform comparably to longer measures for case detection(26,27) and, for some conditions, sufficiently sensitive to change for outcome monitoring.(28-30) The option to administer PROMIS scales using a computerized adaptive testing (CAT) approach allows a precise assessment of a domain with as few as 7 to 8 items and also provides automatic calculation of scores along with normative information pertaining to the US general population.

Although items that focus on a single symptom or problem are generally desirable, bundling more than one symptom or problem into a compound item is occasionally justifiable. One type of compound item uses a few synonyms for the same symptom (e.g., “feeling down, sad, blue, or depressed”; “feeling tired or having little energy”; “feeling your heart pound or race”), since some respondents may identify more with one synonym than another. Some compound items ask about different symptoms but ones that are conceptually or mechanistically related (i.e., emanating from the same organ system or disease), such as “nausea or vomiting”; “constipation or diarrhea”; “pain in your arms, legs or joints”. Here the testing and/or treatment may be the same, so asking separate items is unnecessary from a pragmatic standpoint. A third type of compound item asks about symptoms not so closely related (at times, even opposites) but any one of which fulfills the same diagnostic criterion. For example, the PHQ-9 depression scale has one item about “poor appetite or overeating”, another about “trouble falling or staying asleep or sleeping too much”, and still another about “feeling bad about yourself — or that you are a failure or have let yourself or your family down”.

6. Response options

Dimensions of a symptom or problem commonly assessed by response options include: a) Frequency (how often): e.g., never/rarely/sometimes/often/always; b) Severity/intensity (how much), e.g., not at all/a little bit/somewhat/quite a bit/very much; and c) Impairment, captured by modifiers such as “How much did (symptom) interfere with ...?” or “How much were you bothered (distressed) by (symptom)?” or “How difficult did (symptom) make it to do your...?”(31) A few measures capture multiple dimensions (e.g., the Memorial Symptom Assessment Scale asks about frequency, severity, and impairment for each symptom endorsed by the patient).(32) Most PROs, however, settle on one dimension per item to reduce respondent burden and to simplify scoring. Moreover, it is possible that many respondents recognize the ordinal nature of a response set, regardless of dimensions or words used, understanding that as they go from left to right a symptom is either being ranked as better or worse, and they situate themselves on this ordinal scale accordingly.

The *number* of response options when the choices are words typically range from 3 to 5; it is not clear that many respondents can make finer distinctions than this.(33-35) Likert-type scales sometimes include up to 7 options, although often with only 2 or 3 verbal anchors (for the extremes of the scale and sometimes the middle point). Numerical rating scales may range up to 11 points (e.g., the commonly-used 0 to 10 pain scale). Even if the decision is made to assess only one dimension per symptom, the decision remains of whether to have more than one response set for different symptoms or for different subscales within the measure. For example, the PHQ assesses 5 different types of mental disorders, and does not use the same response set for each disorder. Similarly, the PROMIS profiles use different response sets for different scales, and sometimes more than one response set for subsets of items within the same scale. The advantage of a uniform response set is that respondents do not need to make the cognitive shifts that might be necessary when moving from one response set to another. The disadvantage is that a particular response set may appear more suitable for certain items or subscales, and forcing all items into the same Procrustean bed of responses may be a poor fit.

Selecting the timeframe for an item (e.g., past week, past 2 weeks, past month) is a balancing act between recall accuracy (presumably favored by shorter timeframes) and the benefits of "time-averaging" (so that one bad day or week does not overestimate the magnitude of the problem). Even with longer recall periods, patient grading of a symptom may be overly influenced by its current or recent magnitude.(36) Another factor that may affect the selection of timeframe is the length of time one expects a symptom to improve, either naturalistically or in response to treatment. For example, pain may respond more rapidly to treatment than depression, and physical and role functioning improvement may lag behind symptom improvement.

Deciding on the appropriate intervals between options within a response set is also important. Some options may seem quite close (e.g., "never" and "rarely") but can be appropriate depending upon symptom prevalence, floor and ceiling effects, and other factors. When the distance between items appears small (e.g., one scale has options of "a great deal" and "a very great deal" (37)), it is important to determine if respondents can actually discriminate such subtle differences. Conversely, one should also ask if there are "gaps" between items (e.g., in our experience with the PHQ-9 some respondents desired an option between "not at all" and "several days").

7. Scoring

Simplicity in scoring facilitates clinical uptake. Easy scoring is exemplified by a measure that provides a single summative score of individual items without the complexity of reverse-scored items, transformation of raw scores into standardized scores, or several subscale scores for a single construct.(17) Reverse scoring complicates the simple addition of item scores into a total score and requires respondents to shift the direction of their thinking about responses in moving from one item to another. The rationale that reverse scored items aid in detecting inconsistent responses or response sets (the tendency to check the same response option for most or all of the symptoms) has poor empiric support; indeed, such items may decrease reliability and validity(38-41) and increase the burden for certain

respondents(42,43). Similarly, the need to transform raw scores into a standardized score through a formula or other conversion process also complicates clinical use. Finally, clinicians like easy-to-remember cutpoints whenever possible; for example, 5, 10, and 15 represent thresholds for mild, moderate, and severe levels of symptoms on the PHQ depression, anxiety, and somatization scales.(18)

What is the role of composite scores? As noted earlier, a specific symptom may be assessed along several dimensions in which case the clinician may be provided each score separately, a composite score, or both. For example, some experts believe it is important to measure pain severity and pain interference separately(44), although recent evidence has shown that a single composite pain score is equally responsive.(29,30) Certainly, a single rather than multiple scores for the same symptom is easier to act upon: how does the clinician treat and follow two different numbers for the same symptom, and what action is taken if the numbers are discordant? A second type of composite score combines separate scores for symptom domains that, while conceptually and empirically distinct, commonly overlap, strongly correlate with one another, and may share common treatments. This is the case with two mood symptoms – depression and anxiety – for which it may be valuable to have both separate and composite scores.(45-47) A third type of composite score is for symptoms that, although clearly distinct from one another, commonly cluster, have adverse effects on one another, and may benefit from co-management. For example, the SPADE pentad – sleep, pain, anxiety, depression, and energy (fatigue) – comprises the most prevalent, disabling and undertreated symptom cluster in cancer patients as well in patients with many other medical and mental disorders.(48,49) While there are symptom-specific treatments for each, there are also treatments that work across more than one SPADE symptom(49,50), in which case there may be benefits for clinicians having both individual symptom scores as well as a composite score. Finally, a composite score of multiple seemingly diverse symptoms may represent an underlying construct for which there are specific treatments. For example, patients who report multiple somatic symptoms can be recognized as having somatization or a somatoform disorder which in turn have evidence-based treatments.(51) In such cases, a somatic symptom composite score may even be more useful than multiple individual item-level symptom scores.(18,52,53)

Interpretability of scores is important (6,7), including establishing cutpoints that signify a threshold for case identification or clinical action, determining how much change in a score represents meaningful improvement or worsening, clarifying directionality (i.e., are higher or lower scores worse?), and aiming for scores that can be easily understood by both clinicians and patients in order to facilitate communication and mutual decision-making.

8. Accessibility

Several factors enhance accessibility of a PRO to clinicians and patient populations. First, a measure that is nonproprietary (i.e., available at no cost) is more readily used than one for which there is a charge each time the scale is administered.(54). Second, a measure that is accessible and downloadable from the Internet enhances clinical use. Examples of public domain PROs that are easily available through the Internet include the PROMIS and PHQ scales among others. Third, the more languages into which a PRO has been translated, the

greater is its global reach and applicability to diverse populations. Fourth, the ability to administer the PRO to vulnerable populations is desirable.

Electronic administration of the measure using automated approaches, either through computerized administration in the clinic or by interactive voice recorded (IVR) or web-based data collection outside the clinic, facilitates both accessibility and efficiency of use.(55). The PROMIS Assessment Center (www.promis.org) provides a free web-based tool for data collection. Graphic reports of PROs over time can help direct the clinician to salient issues including changes over time as well as scores that exceed clinical cut offs.(7) Another important step is to incorporate PRO results into the electronic medical records which allows clinicians to not only have the scores "just in time" for decision-making but to also track the trajectory of symptoms over time as treatment is monitored.(56,57)

Strong psychometrics and practical characteristics are two features essential to the use of PROs in clinical settings. Other factors that may influence uptake include using PROs as a metric for assessing quality of care, incentivizing use by payers, incorporating PROs into guidelines for management of particular diseases, and demonstrating to clinicians and patients the utility of measurement-based care for optimizing treatment outcomes. The use of PROs to evaluate and monitor outcomes that principally rely on patient input is gathering momentum(5,58,59) and to accelerate this movement, practical considerations are essential. In the words of the 20th century pragmatic philosopher, Richard Rorty: "It is the vocabulary of practice rather than of theory, of action rather than contemplation, in which one can say something useful about truth."(60) The same might be said about patient-reported outcome measures.

Acknowledgments

Funding:

This work was supported by a Department of Veterans Affairs Health Services Research and Development Merit Review award to Dr. Kroenke (IIR 07-119), VA Career Development Award to Dr. Kean (CDA IK2RX000879), and National Institute of Aging R01 award to Dr. Monahan (R01 AG043465). The sponsor had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

References

1. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010; 63(11):1179–1194. [PubMed: 20685078]
2. James, W. The Pragmatic Method. In: McDermott, JJ., editor. The Writings of William James. Random House; New York: 1967.
3. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010; 63:737–745. [PubMed: 20494804]
4. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered

outcomes and comparative effectiveness research. Qual Life Res. 2013; 22:1889–1905. [PubMed: 23288613]

5. Glasgow RE, Riley WT. Pragmatic measures: what they are and why we need them. Am J Prev Med. 2013; 45(2):237–243. [PubMed: 23867032]
6. Working Group on Health Outcomes for Older Persons with Multiple Chronic Conditions. Universal health outcome measures for older persons with multiple chronic conditions. J Am Geriatr Soc. 2012; 60:2333–2341. [PubMed: 23194184]
7. Snyder CF, Aaronson NK, Chouchair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Qual Life Res. 2012; 21:1305–1314. [PubMed: 22048932]
8. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol. 2014; 67:745–753. [PubMed: 24582946]
9. Hughes EF, Wu AW, Carducci MA, Snyder CF. What can I do? Recommendations for responding to issues identified by patient-reported outcomes assessments used in clinical practice. J Support Oncol. 2012; 10:143–148. [PubMed: 22609239]
10. Klinkman MS. Competing demands in psychosocial care: a model for the identification and treatment of depressive disorders in primary care. Gen Hosp Psychiatry. 1997; 19:98–111. [PubMed: 9097064]
11. Kroenke K. The many C's of primary care. J Gen Intern Med. 2004; 19(6):708–709. [PubMed: 15209611]
12. Arroll B, Goodyear-Smith F, Kerse N, Fishman T, Gunn J. Effect of the addition of a “help” question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. BMJ. 2005; 331(7521):884. [PubMed: 16166106]
13. Kroenke K, Krebs E, Wu J, Bair MJ, Damush T, Chumbler N, et al. Stepped Care to Optimize Pain Care Effectiveness (SCOPE) Trial: study design and sample characteristics. Contemp Clin Trials. 2013; 34:270–281. [PubMed: 23228858]
14. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: a meta-analysis. CMAJ. 2008; 178(8):997–1003. [PubMed: 18390942]
15. Gilbody S, Bower P, Fletcher J, Richards D, Sutton AJ. Collaborative care for depression: a cumulative meta-analysis and review of longer-term outcomes. Arch Intern Med. 2006; 166(21): 2314–2321. [PubMed: 17130383]
16. US Preventive Services Task Force. Screening for depression in adults: US Preventive Services Task Force recommendation statement. Ann Intern Med. 2009; 151:784–792. [PubMed: 19949144]
17. Kroenke K. Enhancing the clinical utility of depression screening. CMAJ. 2012; 184(3):281–282. [PubMed: 22231681]
18. Kroenke K, Spitzer RL, Williams JB, Lowe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry. 2010; 32(4):345–359. [PubMed: 20633738]
19. Richardson LP, McCauley E, Grossman DC, McCarty CA, Richards J, Russo JE, et al. Evaluation of the Patient Health Questionnaire-9 Item for Detecting Major Depression Among Adolescents. Pediatrics. 2010; 126(6):1117–1123. [PubMed: 21041282]
20. Saliba D, DiFilippo S, Edelen MO, Kroenke K, Buchanan J, Streim J. Testing the PHQ-9 interview and observational versions (PHQ-9 OV) for MDS 3.0. J Am Med Dir Assoc. 2012; 13:618–625. [PubMed: 22796361]
21. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. J Gen Intern Med. 2006; 21(6):547–552. [PubMed: 16808734]
22. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res. 2007; 16(Suppl 1):133–141. [PubMed: 17401637]

23. Kobak KA, Taylor LH, Dottl SL, Greist JH, Jefferson JW, Burroughs D, et al. A computer-administered telephone interview to identify mental disorders. JAMA. 1997; 278(11):905–910. [PubMed: 9302242]
24. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU, Ware JE Jr. Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. J Clin Epidemiol. 2014; 67:108–113. [PubMed: 24262772]
25. Monahan PO, Boustani MA, Adler C, Galvin JE, Perkins AJ, Healy P, et al. Practical clinical tool to monitor dementia symptoms: the HABC-Monitor. Clin Interventions Aging. 2012; 7:143–157.
26. Mitchell AJ, Coyne JC. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. Br J Gen Pract. 2007; 57(535): 144–151. [PubMed: 17263931]
27. Hays RD, Reise S, Calderon JL. How much is lost in using single items? J Gen Intern Med. 2012; 27(11):1402–1403. [PubMed: 22878854]
28. Lowe B, Kroenke K, Grafe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). Journal of Psychosomatic Research. 2005; 58(2):163–171. [PubMed: 15820844]
29. Krebs EE, Bair MJ, Wu J, Damush TM, Tu W, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. Med Care. 2010; 48:1007–1014. [PubMed: 20856144]
30. Kroenke K, Theobald D, Wu J, Tu W, Krebs EE. Comparative responsiveness of pain measures in cancer patients. J Pain. 2012; 13(8):764–772. [PubMed: 22800982]
31. Kroenke K. Studying symptoms: sampling and measurement issues. Ann Intern Med. 2001; 134:844–855. [PubMed: 11346320]
32. Portenoy RK, Kornblith AB, Lepore JM, Friedlander-Klar H, Kiyasu E, Sobel K, et al. The Memorial Symptom Assessment Scale: an instrument for the evaluation of symptom prevalence, characteristics and distress. Eur J Cancer. 1994; 30A:1326–1336. [PubMed: 7999421]
33. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica. 2000; 104:1–15. [PubMed: 10769936]
34. Weng LJ. Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educ Psychol Meas. 2004; 64:956–972.
35. Cook KF, Cella D, Boespflug EL, Amtmann D. Is less more? A preliminary investigation of the number of response categories in self-reported pain. Patient Related Outcome Measures. 2010; 1:9–18. [PubMed: 21709756]
36. Smith WB, Safer MA. Effects of present pain level on recall of chronic pain and medication use. Pain. 1993; 55:355–361. [PubMed: 8121697]
37. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Controlled Clin Trials. 1989; 10:407–415. [PubMed: 2691207]
38. Marsh HW. Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? J Personality Soc Psychol. 1996; 70:810–819.
39. Wong N, Rindfleisch A, Burroughs JE. Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. J Consumer Res. 2003; 30:72–91.
40. Stewart TJ, Frye AW. Investigating the use of negatively phrased survey items in medical education settings: common wisdom or common mistake? Acad Med. 2004; 79(Oct suppl):S18–S20. [PubMed: 15383379]
41. Rodebaugh TL, Woods CM, Thissen DM, Heimberg RG, Chambless DL, Rapee RM. More information from fewer questions: the factor structure and item properties of the original and brief fear of negative evaluation scale. Psychological Assessment. 2004; 16:169–181. [PubMed: 15222813]
42. Conrad KJ, Wright BD, McKnight P, McFall M, Fontana A, Rosenheck R. Comparing traditional and Rasch analyses of the Mississippi PTSD Scale: revealing limitations of reverse-scored items. J Applied Measurement. 2004; 5:15–30.

43. Carlson M, Wilcox R, Chou CP, Chang M, Yang F, Blanchard J, et al. Psychometric properties of reverse-scored items on the CES-D in a sample of ethnically diverse older adults. Psychological assessment. 2011; 23(2):558. Psychol Assess 2011; 23:558-562. [PubMed: 21319906]
44. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005; 113(1-2):9–19. [PubMed: 15621359]
45. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res. 2002; 52(2):69–77. [PubMed: 11832252]
46. Kroenke K, Spitzer RL, Williams JBW, Lowe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics. 2009; 50:613–621. [PubMed: 19996233]
47. Cuijpers P, Smits N, Donker T, ten Have M, de Graaf R. Screening for mood and anxiety disorders with the five-item, the three-item, and the two-item Mental Health Inventory. Psychiatry Res. 2009; 168(3):250–255. [PubMed: 19185354]
48. Barsevick AM. The concept of symptom cluster. Semin Oncol Nurs. 2007; 23(2):89–98. [PubMed: 17512435]
49. Kroenke K. A practical and evidence-based approach to common symptoms. Ann Intern Med. 2014 in press.
50. Jackson JL, O'Malley PG, Kroenke K. Antidepressants and cognitive-behavioral therapy for symptom syndromes. CNS Spectr. 2006; 11(3):212–222. [PubMed: 16575378]
51. Kroenke K. Efficacy of treatment for somatoform disorders: a review of randomized controlled trials. Psychosom Med. 2007; 69:881–888. [PubMed: 18040099]
52. Zijlema WL, Stolk RP, Lowe B, Rief W, White PD, Rosmalen JG. How to assess common somatic symptoms in large-scale studies: a systematic review of questionnaires. J Psychosom Res. 2013; 74(6):459–468. [PubMed: 23731742]
53. Gierk B, Kohlmann S, Kroenke K, Spangenberg L, Zenger M, Brahler E, et al. The Somatic Symptom Scale-8 (SSS-8): a brief measure of somatic symptom burden. JAMA Intern Med. 2014; 174(3):399–407. [PubMed: 24276929]
54. Newman JC, Feldman R. Copyright and open access at the bedside. N Engl J Med. 2011; 365(26): 2447–2449. [PubMed: 22204721]
55. Johns SA, Kroenke K, Theobald D, Wu J, Tu W. Telecare management of pain and depression in patients with cancer: patient satisfaction and predictors of use. J Ambulatory Care Management. 2011; 34:126–139.
56. Glasgow RE, Kaplan RM, Ockene JK, Fisher EB, Emmons KM. Patient-reported measures of psychosocial issues and health behavior should be added to electronic health records. Health Aff. 2012; 31:497–504.
57. Wu AW, Kharrazi H, Boulware L, Snyder CF. Measure once, cut twice-adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. J Clin Epidemiol. 2013; 8:S12–S20. [PubMed: 23849145]
58. Harding KJ, Rush AJ, Arbuckle M, Trivedi MH, Pincus HA. Measurement-based care in psychiatric practice: a policy framework for implementation. J Clin Psychiatry. 2011; 72(8):1136–1143. [PubMed: 21295000]
59. Basch E, Torda P, Adams K. Standards for patient-reported outcome-based performance measures. JAMA. 2013; 310:139–140. [PubMed: 23839744]
60. Rorty, R. Consequences of Pragmatism. University of Minnesota Press; Minneapolis, MN: 1982.

What is new?

Key points

- Clinical uptake of patient-reported outcome (PRO) measures requires pragmatic as well as psychometric considerations.
- Eight pragmatic characteristics include actionability, setting-appropriateness, universality, self-administration, item features, response options, scoring, and accessibility.
- Examples from the literature as well as public domain PROs such as the Patient Health Questionnaire (PHQ) and Patient-Reported Outcome Information System (PROMIS) scales as well as other PROs exemplify these pragmatic considerations

Table 1

Characteristics of a Practical Patient-Reported Outcome Measure

	Characteristic	Comment
1.	Actionable	Do scores guide diagnostic or therapeutic action/decision-making?
2.	Setting-appropriate	Is the target audience: a) primary care outpatients; b) specialty clinic outpatients; c) hospitalized patients; d) extended care/other settings?
3.	Universal	Can measure screen, diagnose, assess severity, and monitor therapy? Can measure be used in multiple different diseases and conditions?
a	• Multi-purpose	
b	• Cross-cutting	
4.	Self-administered	Is completion of the measure by the patient alone reliable and valid?
5.	Items	Is the measure sufficiently brief? Are items simple (1 symptom) or compound (≥ 2 related symptoms)?
a	• Number	
b	• Bundling	
6.	Response options	What is being assessed – frequency, severity, impairment, other? How many response options per item (3, 4, 5, or more)? Is there one response set for all items (vs. varying response sets)? What is the optimal recall period (days, weeks, months, or other)? Is there appropriate spacing between options? Are there gaps?
a	• Dimension	
b	• Number	
c	• Uniformity	
d	• Time frame	
e	• Intervals	
7.	Scoring	Can reverse scored items or score transformation be avoided? Are there subscale scores, a composite score, or both? What are meaningful clinical outpoints and changes over time?
a	• Simplicity	
b	• Number of scores	
c	• Interpretability	
8.	Accessibility	Is the measure public domain (i.e., available at no cost)? Is the measure easy to access through the Internet? Has the measure been translated into multiple languages? Can the measure be completed by those with low literacy, disabilities, or impaired capacity (including proxy completion when necessary)? Is the measure captured through automation and in electronic records?
a	• Nonproprietary	
b	• Downloadable	
c	• Translations	
d	• Vulnerable groups	
e	• Electronic	

Table 2

Comparison of Recommended Practical Characteristics for Patient-Reported Outcome Measures *

Group	Methodological Approach †	Practical Characteristics *								
		Actionable	Setting Appropriate	Universal	Self-Administered	Item Issues	Response Options	Scoring	Accessibility	Additional Practical Characteristics
Kroenke et al.	Authors' Perspective	X	X	X	X	X	X	X	X	
Glasgow et al [5]	Authors' Perspective	X	X	X		X			X	• important to stakeholders; • useful to assess care of patient groups (quality improvement) • unlikely to cause harm/liability to practice
Snyder et al [7]	Expert Consensus	X	X		X	X	X	X	X	• mode of administration; • frequency of administration; • static forms vs. dynamic (i.e., computerized item-bank) • useful to assess care of patient groups (quality improvement) • method for presenting scores
NIA Work Group [6]	Expert Consensus	X	X	X	X			X	X	• proxy-administered version; • interpretable by patients

* Various groups may have used a different term for a particular characteristic or mentioned different points related to a characteristic

† All authors/groups used a selected (rather than systematic) literature review