Francis *et al. Systematic Reviews* (2016) 5:129
DOI 10.1186/s13643-016-0307-4

Systematic Reviews

METHODOLOGY Open Access

Checklist to operationalize measurement characteristics of patient-reported outcome measures

David O. Francis[1,2,3*], Melissa L. McPheeters[3,4,5], Meaghan Noud[1], David F. Penson[2,4,6,7] and Irene D. Feurer[2,8]

Abstract

Background: The purpose of this study was to advance a checklist of evaluative criteria designed to assess patient-reported outcome (PRO) measures' developmental measurement properties and applicability, which can be used by systematic reviewers, researchers, and clinicians with a varied range of expertise in psychometric measure development methodology.

Methods: A directed literature search was performed to identify original studies, textbooks, consensus guidelines, and published reports that propose criteria for assessing the quality of PRO measures. Recommendations from these sources were iteratively distilled into a checklist of key attributes. Preliminary items underwent evaluation through 24 cognitive interviews with clinicians and quantitative researchers. Six measurement theory methodological novices independently applied the final checklist to assess six PRO measures encompassing a variety of methods, applications, and clinical constructs. Agreement between novice and expert scores was assessed.

Results: The distillation process yielded an 18-item checklist with six domains: (1) conceptual model, (2) content validity, (3) reliability, (4) construct validity, (5) scoring and interpretation, and (6) respondent burden and presentation. With minimal instruction, good agreement in checklist item ratings was achieved between quantitative researchers with expertise in measurement theory and less experienced clinicians (mean kappa 0.70; range 0.66–0.87).

Conclusions: We present a simplified checklist that can help guide systematic reviewers, researchers, and clinicians with varied measurement theory expertise to evaluate the strengths and weakness of candidate PRO measures' developmental properties and the appropriateness for specific applications.

Abbreviations: PRO, Patient-reported outcome; FDA, Food and Drug Administration; HRQOL, Health-related quality of life; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; MID, Minimally important difference; MPH, Masters of Public Health; PhD, Doctor of Philosophy; DrPH, Doctor of Public Health

Background

Improved health expectations have led to a shift away from viewing health in terms of survival toward defining as freedom from disease, followed by concentration on an individual's ability to perform daily activities, and more recently to an emphasis on themes of well-being and quality of life [1–4]. Concomitant to the evolving conception of population health has been a transition from reliance on clinically focused end points without direct input from patients [5, 6] to increased emphasis on patient-centered outcome research and comparative effectiveness research [7]. As such, patients, families, and clinicians are increasingly faced with complex choices and ambiguous information when addressing health and healthcare needs.

* Correspondence: david.o.francis@vanderbilt.edu
[1]Department of Otolaryngology, Vanderbilt University Medical Center, Medical Center East, Suite 7302, 1215, 21st Avenue South, Nashville, TN 37212, USA
[2]Center for Surgical Quality and Outcomes Research, Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville 37232, TN, USA
Full list of author information is available at the end of the article

It is important to differentiate between patient-centered *data* and patient-centered *outcomes*. Data are information deriving directly from patients, and outcomes are end points that matter to patients [6, 7]. A National Institutes of Health/Food and Drug Administration (FDA) working group identified three categories: *feeling*, *function*, and *survival* as primary patient-centered outcomes to be focused on and incorporated into all clinical trials proposing novel interventions, devices, or pharmaceuticals that aim for FDA approval [5]. A significant challenge in patient-centered outcome research and comparative effectiveness research is how best to identify and use patient-centered outcomes that measure effectiveness, facilitate decision-making, and inform health policy [8]. Patient-reported outcome (PRO) measures are now commonly used in this capacity and are defined as “any report on the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else” [8, 9].

Nomenclature in this field is nuanced and PROs, PRO measures, and health-related quality of life (HRQOL) are often used interchangeably [10, 11]. *Health-related quality of life* is “the value assigned to duration of life as modified by impairments, functional status, perceptions, and social opportunities that are influenced by disease, injury, treatment, or policy” [11–15]. In distinction, *PROs* provide reports *directly* from patients about health, quality of life, or functional status related to the healthcare or treatment they have received [6, 16], and *PRO measures* are designed to measure and report PRO constructs [6, 17]. We have chosen to use the term “PRO measure” heretofore to encompass the various types of health-related instruments including HRQOL, recognizing that others may prefer other terms [10, 16, 18]. Our rationale is that these types of instruments span a diverse gamut that include symptom indices [19, 20], general [21] and condition-specific HRQOL [22, 23], utilities [24, 25], well-being [26, 27], or social health [28] or can focus on latent constructs such as self-efficacy [29] and willingness to change [30, 31].

Patient-reported outcome measures address the need for patient-centered data and are now used in diverse clinical, research, and policy pursuits [32]. Greater emphasis on patient-centered care has resulted in instrument proliferation [33]. However, their developmental rigor and intended application vary widely [34], and this variation is likely to be reflected in systematic reviews. For instance, these instruments can be used as outcomes for group-level analyses in clinical trials and observational studies [35], but are also used to track within-person change over time [36], for group-level quality improvement initiatives to provide information for report cards [37], and as health surveys to monitor population health [38, 39]. In practice, a specific measure may be used in any or all these applications.

Patient-reported outcome measures have origins in various measurement theory-related disciplines including psychometrics [40], clinimetrics [41], and econometrics [4]. There is considerable overlap in approach between these disciplines, and collectively, they strengthen quantitative design methodologies. The common core principles of measure development are multifaceted and sometimes complex. Identifying the appropriate PRO measure for a particular purpose requires nuanced understanding of a candidate measure’s underlying conceptual model and its measurement properties [16]. Most clinicians, researchers, and patient advocates are not experts in the technical methods used to develop and validate these tools and may, understandably, presume similar performance among published PRO measures that address a particular construct. This is problematic since nearly all published tools purport some degree of these attributes, most often as forms of reliability or validity [34].

To address this issue, increased attention has been directed toward understanding what defines adequacy among PRO measures [5, 6, 10, 34, 42, 43]. This is directly relevant to systematic reviewers choosing to incorporate PRO measures as outcomes for their reviews. Current expert panel recommendations and proposed criteria on this topic have substantial homology, but differences do exist [6, 10, 34, 42–44]. Some advanced criteria are not easily understood, and others are rigorously prescriptive, tending to render most instruments inadequate in several respects. These concerns have contributed to disparate quality among systematic reviews of PRO measures and have the potential to mislead researchers into reliance on inappropriate or suboptimal instruments for a given purpose [10, 45–47]. For example, measurement bias in estimation of treatment effects can occur due to lack of conceptual equivalence between PRO measures [47].

An important and rigorous effort to aid researchers in the selection of appropriate PRO measures, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN), was devised between 2006 and 2010 by an expert panel with diverse backgrounds (e.g., clinical medicine, biostatistics, psychology, epidemiology) [11, 16, 44, 46, 48]. Consensus was achieved as to measurement properties that should be assessed and on criteria for acceptable measurement [11, 16, 44]. Three overarching measurement domains were agreed upon: reliability, validity, and responsiveness. The product of this important work was a detailed algorithm for each identified domain. COSMIN remains the standard in the assessment of patient-reported outcome measures. However, its complexity (i.e., 119 items over 10 categories) may limit its utility for a systematic reviewer, researcher, or clinician who may not have expertise in measurement theory. Furthermore, its stated use is for evaluative instruments designed for applications to measure change over time. It may not apply for discriminative instruments, those used for predictive purposes, or healthcare-related measures used to measure satisfaction with care or adherence [16].

A simplified methodology that incorporates the critical features highlighted in COSMIN and other pertinent literature would be helpful to enable systematic reviewers, researchers, and clinicians to assess developmental characteristics and usefulness of a wide variety of PRO measures. In addition, its usefulness would be enhanced by the inclusion of practical aspects of PRO measures not consistently addressed in other criteria [6]. Thus, our study aimed to (1) advance a set of simplified criteria, in the form of a checklist, that can aid in systematically assessing the measurement properties and usefulness of PRO measures for particular circumstances and 2) demonstrate the checklist's user-friendliness by determining the inter-rater reliability of its scoring between clinicians/researchers with and without expertise in empirical instrument developmental methods. The resultant checklist is intended as a guide for systematic reviewers, researchers, and clinicians with diverse measurement theory expertise to aid in identifying the strengths, weaknesses, and applicability of candidate PRO measures.

Methods

A review of the literature was performed to identify recommendations for evaluating PRO measures. The directed search enabled the compilation of PRO measures' developmental recommendations from a wide variety of sources including the FDA [5, 6], the Scientific Advisory Committee of the Medical Outcomes Trust [43, 49], COSMIN [11, 16, 44, 46], Agency for Healthcare Research and Quality [10], American Psychological Association [50, 51], measurement theory textbooks [40, 52–54], and individual studies via a PubMed search for evaluative criteria germane to PRO measures, health-related quality of life, and related terminology. This study did not involve data collection from or about human subjects and was therefore exempt from IRB review.

Two investigators (DOF, IDF) analyzed and synthesized these recommendations and iteratively distilled them into initial criteria. Attributes considered fundamental were (1) conceptual model, (2) content validity, (3) reliability, (4) construct validity, (5) scoring and interpretation, and (6) respondent burden and presentation. Founded in psychometrics (e.g., classical test and item response theories) [40, 43, 55, 56] and clinimetrics [57], the core qualities outlined below encompass the theoretical underpinnings of a PRO measure and the developmental characteristics necessary to ensure its overall usefulness.

1. *Conceptual model* provides a rationale for and description of the concepts and the populations that a measure is intended to assess [8, 43, 58, 59]. The *concept* is the specific measurement goal and should be explicitly stated in the development process. Conceptual models are developed by outlining hypothesized and potential concepts and relationships and by determining the target population and model's application [6, 58, 60]. In assessing its adequacy, a candidate measure's original development should be examined to determine if it is likely to capture the intended effect [10]. Whether multiple domains or subscales are expected should be clearly inherent to or directly pre-specified within the conceptual framework [8, 43, 61]. Ninety percent of International Society for Quality of Life Research survey respondents endorsed that PRO measures should have documentation defining the construct and describing the measure's application in the intended population [5, 8, 43].
2. *Content validity* refers to evidence that a PRO measure's domain(s) is appropriate for its intended use in both relevance and comprehensiveness [10, 43, 46, 61, 62]. No formal statistical test exists to evaluate content validity. Instead, assessment is done through applying qualitative criteria. Specifically, items (i.e., questions) and conceptual domains (e.g., subscales) should be relevant to target patients' concerns. Thus, developers should obtain input from the target population to optimize item relevance and clarity, ideally, through qualitative focus groups and cognitive interviews [5, 61]. In brief, cognitive interviews are a qualitative research tool used to determine whether respondents understand included concepts and items in the way that PRO measure developers intend. These interactive "field-test" interviews allow developers to better understand how respondents interpret candidate questions [6]. Similarly, content experts should participate in PRO measure development with emphasis on evaluating the relevance of items for the construct and for the respondent population [43, 46, 61, 62], and there should be a thorough description of how items were elicited, selected, and developed [5].
3. *Reliability* is the degree to which scores are free from random (measurement) error [11, 43]. Several forms exist. *Internal consistency reliability*, the degree to which segments of a test (e.g., split halves, individual items) are associated with each other [56], reflects precision at a single time point [43]. It is based on correlation of scores between different items within

   the PRO measure, thus assessing whether items proposed to measure the same general construct or domain are statistically related. *Test-retest reliability* refers to the reproducibility or stability of scores over two administrations, typically in close temporal proximity, among respondents who are assumed not to have changed on the relevant domains [43, 56]. Traditionally cited minimum levels for reliability coefficients are 0.70 for group-level comparisons and 0.90 to 0.95 for individual comparisons [8, 43]. Coefficients indicate the ratio of true score variance to observed score variance. These thresholds are important to establish the reliability of an instrument. However, some argue that establishing absolute thresholds for interpreting coefficients may be overly prescriptive [8, 53]. Therefore, reliability estimates lower than the convention cited above should be justified in the context of the proposed PRO measure's intended application, its sample size, and the reliability statistic used [63].

4. *Construct validity* refers to whether a test measures theoretic intended constructs or traits [40, 43, 56], and it directly affects the appropriateness of measurement-based inferences. Evidence of construct validity can derive from empirical demonstrations of dimensionality [5, 55]. A variety of latent variable modeling techniques such as factor analysis are available to evaluate and provide evidence of dimensionality, and these methods should be used and reported when subscales or domains are proposed or expected. Factor analysis (and related latent variable methods) is, in general, a data reduction method intended to mathematically represent a large number of differentially related questions (i.e., items) by a smaller number of latent dimensions or "factors." A factor is a mathematical representation of a collection of variables that are statistically related to one another, which differs conceptually from other factors [53, 55]. Generally speaking, factor analysis methods such as common factor analysis, principal components analysis, and bi-factor analysis are important in both classical and item response theory-based instrument development processes [55].

   Responsiveness to change, which is also known as longitudinal construct validity [64], can be considered an aspect of validity [65] or as a separate dimension [11]. It is the extent to which a PRO measure detects meaningful change over time when it is known to have occurred [8, 43, 66]. Most, but not all, instruments have a stated goal of measuring change over time. Thus, this property is not applicable to PRO measures intended specifically for cross-sectional study designs. If a measure is not intended to measure change (e.g., screening test), this point should be specified in the conceptual model. Responsiveness requires demonstrable test-retest reliability *and* the ability to detect an expected change (e.g., after intervention) in the intended population [43, 66]. Absence of either element limits the confidence that measured differences in scores represent an actual change rather than measurement error.

   Responsiveness to change can be measured using two approaches: distribution- or anchor-based methods. Distribution-based methods are based on either within-group change over time or between-group comparisons. Such approaches are characterized by an effect size, standard response mean, or as other measures that account for actual change related to random error (e.g., standard error of measurement) [10]. Anchor-based methods quantify differences by examining the relationship between the PRO measure score and an independent measure (anchor) that could be patient-based, physician-based, or an alternate external assessment of construct severity [67–69]. Both methodologies necessarily incorporate both expected change and test-retest reliability in their calculation. Candidate PRO measures' responsiveness characteristics are particularly relevant for systematic reviewers aiming to compare effectiveness of interventions.

   Another form of construct validity is the degree to which PRO measure scores correlate with other questionnaires that evaluate the same construct or with related clinical indicators (e.g., pulmonary function tests) [43, 56]. This is sometimes referred to as "convergent validity." A priori hypotheses about expected associations between a PRO measure and similar or dissimilar measures should be documented [8, 43]. A closely related concept, called "known groups" or *divergent validity*, requires the PRO measure to differentiate between groups that past empirical evidence has shown to be different. These types of validity have also been classified under the auspices of hypotheses testing [46].

   It is rarely possible to establish a PRO measure's criterion validity because, in the majority of cases, no "gold standard" exists to measure the targeted construct [8]. It is, however, a pertinent parameter in questionnaires designed to be predictive of a certain state (*predictive validity*). For example, self-rated health has been shown to predict mortality [70]; thus, predictive validity can be considered a form of *criterion-related validity*. A clear distinction needs to be made between predictive and longitudinal validity (responsiveness). The former refers to the ability of a "baseline" score (e.g., test result) to predict some future event [53] and is reflected by that association. It does not imply a measure's ability to distinguish change between initial and follow-up assessments.

5. *Scoring and interpretation. Interpretability* is the degree to which the meaning of scores is easily understood [5, 8, 43, 71]. This requires that a *scoring system* be clearly described and that some form of *scaling* exists to indicate what different scores mean. A scoring system defines how to compute scores, whether as a total score or subscales, on the basis of empirical evidence (e.g., a principal component structure supporting a particular number of subscales). Scaling properties depend on the context of the measurement instrument. Total score and item-level scaling are often used and several methodologies exist, including those from classical test theory (e.g., standard error of true scores) [55, 56] and item response theory (e.g., Rasch modeling) [55, 56]. Empirically based scaling allows end users to readily interpret scores, as does considering the availability of relevant population-level or condition-specific normative data or "norms," which permit referencing scores to appropriate standards.
   It is important to understand what represents a minimally important difference and to have the ability to differentiate degrees of difference (e.g., severity) for the construct [72]. Minimally important difference (MID) is defined as "the smallest difference in score in the outcome of interest that informed patients or proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in management" [73]. In brief, MID can be established using distribution- or anchor-based approaches. The anchor should be independently interpretable, and there must be reasonable correlation between the PRO measure score and anchor [72, 74]. The distribution-based method uses the magnitude of change compared to the variability in scores (e.g., effect size). A salient argument is also made that the term "patient-important" is more appropriate than "clinically important" to emphasize the patient-centrism of these outcomes and the goals of directed interventions [18, 75]. The meaningfulness of differences should ideally be based on what patients consider a minimally important, small, moderate, and large difference [76]. Incorporating patients' perspective on what constitutes a difference strengthens the clinical usefulness of the PRO measure. Without this information, it can be difficult to contextualize longitudinal or cross-sectional outcomes and understand if the magnitude of change is important.
   Finally, an often overlooked aspect of scoring and interpretability is an explicit plan for *managing and/or interpreting missing responses* [77], which are common in the practical use of PRO measures [78]. Missing item data introduces error in individual score computation. Data that are missing in a systematic manner may introduce bias into group- and population-level analyses. Several methods exist to manage missing responses and data, and instructions regarding how to manage missing responses are important. Without them, the user is often left to score only those surveys for which responses are complete.
6. *Respondent burden and presentation. Burden* refers to the time, effort, and other demands placed on respondents or those administering the instrument [43, 71]. Acceptable burden in the context of the number of items in and the time necessary to complete a PRO measure is somewhat subjective and depends on the measure's intended use. Lengthy measures might be considered reasonable in a research setting but overly burdensome if administered during a busy clinic. These issues should be explicitly considered, as overly burdensome PRO measures can limit their applicability and practical adoption into studies [79]. The length of a PRO measure should be contextually appropriate [71].
   Another consideration of burden and presentation is the *literacy level* required to understand and complete the measure [80]. Most experts recommend that items be at the sixth grade reading level or lower; however, this criterion should be contextualized to the intended target population [8] and it should be justified. Finally, a PRO measure's items and their presentation should be available to be viewed or accessed by persons considering incorporating its use into practice [71]. Without this level of transparency, it is difficult to fully evaluate a prospective instrument's appropriateness for a particular application.

Cognitive interviews

Our goal of distilling key criteria into a checklist was to provide guidance on how to systematically evaluate candidate PRO measures' developmental characteristics and usefulness for a particular purpose. The intended audience for the proposed criteria is systematic reviewers, researchers, and clinicians with varied expertise in PRO measure development and application. Thus, the initial criteria checklist was reviewed by a group of 12 clinicians (medical students [*n* = 3], physicians [*n* = 9]) and 12 investigators with expertise in survey-based quantitative methods (MPH [*n* = 6], PhD/DrPH [*n* = 6]). Each participant was asked to review and comment on the clarity, accuracy, completeness, and user-friendliness of the criteria. Study personnel asked respondents directed follow-up questions to foster discussion and further clarification of concerns. Comments were used to improve clarity, readability, accuracy, and completeness and to establish the revised final criteria checklist (Fig. 1).

Inter-rater reliability of the checklist

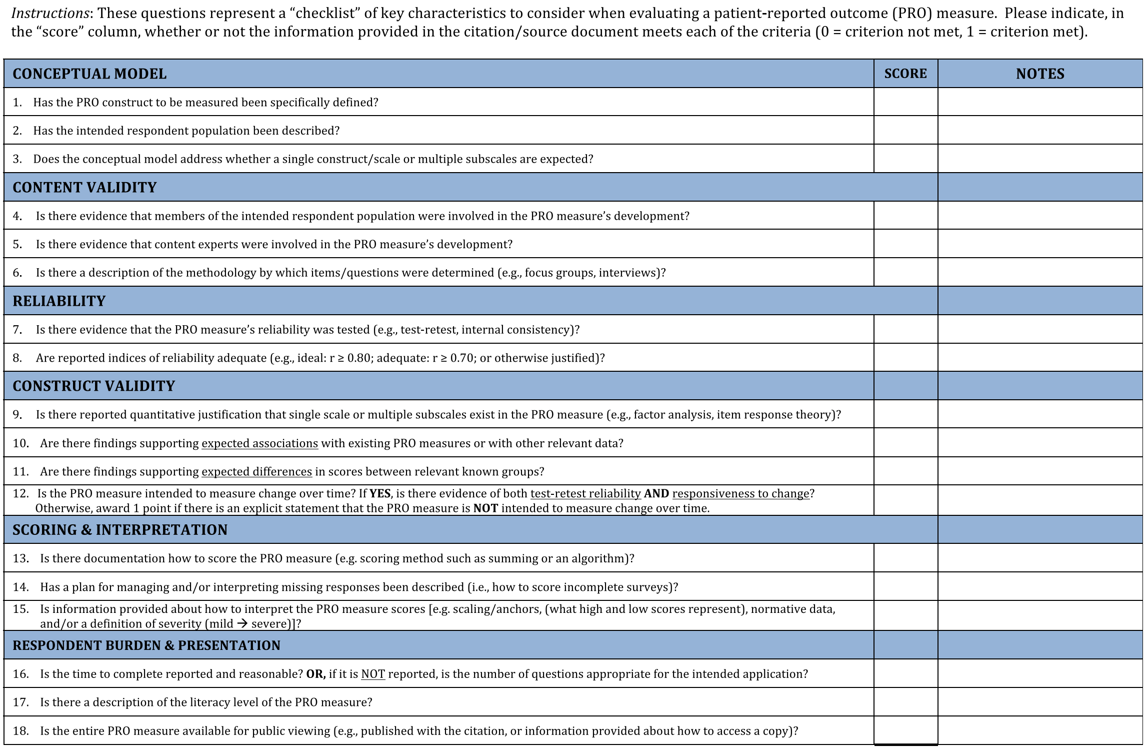
Two investigators (DOF, IDF) used the checklist to assess six pre-specified PRO measures encompassing a variety of methods and applications related to voice and swallowing disorders [81–86]. Two measures were designed to measure handicap (VHI, VHI-10) [81, 82], and one each was designed to measure health-related quality of life (V-RQOL) [85], coping (VDCQ) [84], and activity and participation (VAPP) [87] associated with voice disorders. Another measure developed using item response theory techniques focused on health-related quality of life among patients with achalasia [86]. Discordances were resolved with a modified Delphi technique, and agreed upon criterion-level decisions and tallies provided reference values for each measure. A group of six clinicians without expertise in measurement theory graded the six PRO measures, and their agreement with reference "scores" was summarized as the kappa statistic. An a priori threshold for kappa was set at greater than 0.50 for each PRO measures to demonstrate at least moderate agreement. A stepwise process was used. Participating clinicians were first provided with the checklist (Fig. 1) and brief written descriptions of concepts being evaluated (Additional file 1: Figure S1). Each independently scored the PRO measures, and if kappa scores were inadequate, participants were provided 15 min of individualized education on the concepts followed by rescoring as necessary. This process provided more in-depth information and detailed feedback regarding parameters included in the criteria.

Results

Cognitive interviews

The cognitive interviews highlighted that several respondents were concerned that the checklist mentioned technical detail or sophisticated concepts that the average user would not be familiar with (e.g., factor analysis, item response theory). Responding to these concerns, an addendum was created and appended (Additional file 1: Figure S1).

Respondents expressed concern that some criteria did not have strict benchmarks for decision-making. An example is "has the PRO construct been specifically defined?" Supporting documentation was clarified to note that these criteria are necessarily general (and somewhat vague) due to their inherent subjectivity and absence of specific standards. One respondent questioned whether the target population's demographic or clinical characteristics should be defined, and several recommended



Instructions: These questions represent a "checklist" of key characteristics to consider when evaluating a patient-reported outcome (PRO) measure. Please indicate, in the "score" column, whether or not the information provided in the citation/source document meets each of the criteria (0 = criterion not met, 1 = criterion met).

CONCEPTUAL MODEL	SCORE	NOTES
1. Has the PRO construct to be measured been specifically defined?		
2. Has the intended respondent population been described?		
3. Does the conceptual model address whether a single construct/scale or multiple subscales are expected?		
CONTENT VALIDITY		
4. Is there evidence that members of the intended respondent population were involved in the PRO measure's development?		
5. Is there evidence that content experts were involved in the PRO measure's development?		
6. Is there a description of the methodology by which items/questions were determined (e.g., focus groups, interviews)?		
RELIABILITY		
7. Is there evidence that the PRO measure's reliability was tested (e.g., test-retest, internal consistency)?		
8. Are reported indices of reliability adequate (e.g., ideal: r ≥ 0.80; adequate: r ≥ 0.70; or otherwise justified)?		
CONSTRUCT VALIDITY		
9. Is there reported quantitative justification that single scale or multiple subscales exist in the PRO measure (e.g., factor analysis, item response theory)?		
10. Are there findings supporting expected associations with existing PRO measures or with other relevant data?		
11. Are there findings supporting expected differences in scores between relevant known groups?		
12. Is the PRO measure intended to measure change over time? If **YES**, is there evidence of both test-retest reliability **AND** responsiveness to change? Otherwise, award 1 point if there is an explicit statement that the PRO measure is **NOT** intended to measure change over time.		
SCORING & INTERPRETATION		
13. Is there documentation how to score the PRO measure (e.g. scoring method such as summing or an algorithm)?		
14. Has a plan for managing and/or interpreting missing responses been described (i.e., how to score incomplete surveys)?		
15. Is information provided about how to interpret the PRO measure scores [e.g. scaling/anchors, (what high and low scores represent), normative data, and/or a definition of severity (mild → severe)]?		
RESPONDENT BURDEN & PRESENTATION		
16. Is the time to complete reported and reasonable? **OR,** if it is NOT reported, is the number of questions appropriate for the intended application?		
17. Is there a description of the literacy level of the PRO measure?		
18. Is the entire PRO measure available for public viewing (e.g., published with the citation, or information provided about how to access a copy)?		

Fig. 1 Checklist to operationalize developmental characteristics and applicability of patient-reported outcome measures

simplifying grammar and sentence structure. They unanimously questioned the propriety of summing the criteria into a total score and felt that the individual criteria presented did not warrant uniform weights.

Some respondents also recommended removing strict thresholds for interpreting reliability. Despite this recommendation, we opted to include them because they represent important, accepted conventions, especially since less experienced users need some guidance regarding interpretation. Some respondents felt it would be helpful to parenthetically list types of reliability that should be tested (e.g., test-retest reliability, internal consistency), and some questioned whether testing dimensionality through factor analysis or other quantitative approaches should be classified as a component of reliability rather than validity. While the characteristics of scales and the items comprising them can be assessed for their internal consistency reliability, we opted to present this concept in the construct validity section, with the rationale that empirically identified dimensions should reflect the conceptual domains represented by the PRO measure.

Another characteristic that proved difficult for some respondents related to responsiveness. This question required that the PRO measure demonstrate both test-retest reliability *and* evidence of responsiveness to change. The rationale for the prerequisite of test-retest reliability was that if a PRO measure has not shown stability then evidence of responsiveness cannot be proven. Several reviewers suggested splitting this question so that it only takes into account responsiveness to change. Others recommended using the term “changes over time” rather than responsiveness or longitudinal validity.

Several persons recognized the subjectivity of asking reviewers to assess whether the PRO measure length was “reasonable.” Initially, an example length of 10 items was included if no mention of burden was mentioned. However, most respondents felt that was too prescriptive and that longer measures were not overly burdensome in specific circumstances. There was also question whether the ability to access the entire PRO measure really mattered. All of these issues were carefully considered, and many suggestions were incorporated in the final criteria.

Proposed checklist

Shown in Fig. 1 is the proposed criteria checklist for assessing the development characteristics and utility of PRO measures. Eighteen characteristics are to be scored dichotomously (present/absent) in six general domains: conceptual model (three items), content validity (three items), reliability (two items), construct validity (four items), scoring and interpretability (three items), and respondent burden and presentation (three items). On the basis of feedback from cognitive interviews and in consideration of the stage of the instrument’s development, individual characteristics and domains were not weighted. The final criteria are referred to as a checklist and are intended as a guide when selecting or evaluating PRO measure.

Agreement between novice users and reference scores

All six participating clinicians independently scored the same six individual PRO measures (*n* = 36). Overall, the mean clinician kappa for the first iteration (written instruction only) was 0.54 (range 0.35–0.63) with 21/36 reviews meeting the a priori criterion of kappa greater than 0.50. One clinician met the criterion on all six PRO measures on the first attempt (Table 1). Two participants met the threshold for 5/6, and one each met the threshold for 3/6, 2/6, and 1/6 of tested PRO measures.

The five remaining participants each received brief education on concepts and rescored the measures for which agreement was below the criterion (*n* = 15). In the second iteration, four of five participants achieved adequate agreement on all measures (Table 1). One required a second educational session, followed by rescoring, and thereafter achieved adequate agreement on all PRO measures. The final mean kappa statistic for the clinicians was 0.70 (range 0.66–0.87; Table 1).

Discussion

We distilled existing consensus criteria into a checklist that can be readily employed in systematic reviews that aim to assess PRO measures’ developmental properties. This checklist provides end users a means to evaluate the appropriateness of PRO measures prior to applying them

Table 1 Interventions and novice reviewer agreement with reference scores

Reviewer	Written	Kappa (range)	1st teaching	Kappa (range)	2nd teaching	Kappa (range)
1	X	0.63 (0.50–0.82)				
2	X	0.59 (0.25–0.77)	X	0.66 (0.56–0.77)		
3	X	0.34 (0.02–0.61)	X	0.66 (0.56–0.82)		
4	X	0.51 (0.27–1.00)	X	0.72 (0.54–1.00)		
5	X	0.64 (0.49–1.00)	X	0.66 (0.51–1.00)		
6	X	0.54 (0.09–0.85)	X	0.65 (0.33–1.00)	X	0.87 (0.68–1.00)

X indicates that this type of instruction was provided to the reviewer. First provided was written instructions followed by in-person teaching (teaching was ≤15 min). Kappa scores: mean (range) for research or clinical purposes. The checklist's strength is the demonstration that, with minimal instruction, systematic reviewers, researchers, and clinicians with limited PRO measure methodological expertise can apply it with ratings that correlate highly with experts in instrument development methodology.

There are long-standing discussions about what constitutes quality among survey and test instruments that even occur in the fields of psychology and education, where measurement theory was initially developed and promulgated. An initial consensus statement in 1954 identified the core qualities of survey development as dissemination, interpretation, validity, reliability, administration and scoring, scaling, and norms [50]. Social scientists, statisticians, and health outcome researchers have refined and advanced these developmental methodologies; however, the same principles first described still pervade consensus statements and expert opinion in the fields of education, social science, and healthcare.

Incorporation of PRO measures' developmental methodology in healthcare has evolved rapidly with the emergence of comparative effectiveness research and patient-centered outcome research. Feinstein aptly described the foundation of this important work stating that “assessment of health status is important because improvements in symptoms, other clinical problems, and functional capacity are usually the main goals of patients in seeking care” [88]. Patient-reported outcome measures are increasingly used to better understand the perspectives of and to measure concepts that matter to the patient [5]. Methodological experts in PRO measures and survey design have disseminated several consensus statements to guide appropriate development and implementation of these measures [5, 8, 10, 43, 89]. Use of poorly developed PRO measures or those designed for a purpose that differs from their use can have significant implications and lead to distorted, inaccurate, or equivocal findings [5, 47]. Measures should be chosen based on relevance and their track record in the context of the proposed study [10]. Therefore, it is incumbent upon researchers and other end users to carefully consider a measure's properties and weigh its strengths and potential weaknesses before implementing it in practice, clinical trials, quality improvement initiatives, or population-level studies.

Simplified access to evaluation criteria should encourage easier and more careful vetting of candidate PRO measures by potential end users. It can be applied to evaluate a specific instrument's characteristics or in the performance of systematic reviews of PRO measures' developmental properties. The complexity and prescriptiveness of prior consensus guidelines on PRO measure development may limit their practical application by systematic reviewers, researchers, and clinician end users who are not expert in survey design and measurement theory. To overcome this issue, we have advanced a simple checklist for evaluating the adequacy of any survey or PRO measure. It cannot be over emphasized that its contents are not intended to replace prior consensus statements on this topic. Instead, it aims to distill and harmonize homologous concepts that have been widely recognized in published expert consensus statements.

Considerations and limitations

Our proposed checklist is not exhaustive. Psychometric and clinimetric PRO measures' development principles are often complex, conceptually overlapping, and evolving [90]. It is not possible to accommodate and incorporate all parameters and circumstantial caveats within simple criteria. One example is administrative burden (e.g., personnel time needed to help patients complete questions), which can affect the ease of application of a particular PRO measure and was not explicitly addressed in the present checklist. Further, it is important to recognize that the fundamental principles of survey development exist on a spectrum, are often interchangeable, and are not necessarily discrete concepts. An example is responsiveness, which has been categorized as an aspect of validity [64, 65] but also as its own domain [11]. Additionally, because each checklist characteristic was derived from broadly accepted core concepts in survey methodology and measurement theory that by their nature are not necessarily expected to correlate with each other, the utility of latent variable methods such as factor analysis are not applicable at this stage.

The relative importance of a specific measurement property may vary substantially with the purpose and context of a PRO measure's use. As such, we do not recommend a total score for this tool since this implies each item should be weighted equally. Our analysis of inter-rater reliability of ratings between novice and more experienced practitioners of measurement theory was not intended to provide rigorous evidence of the checklist's completeness. Instead, this preliminary analysis was performed to show that this simple checklist was easy to apply and reliable even among those with little expertise the field. The proposed system is designed to serve as a guide to understand the strengths and weaknesses and applicability of any particular survey or PRO measure.

Conclusions

Systematic reviewers, researchers, and clinicians who are considering using a particular PRO measure as an outcome in the performance of or evaluating clinical trial results need to be able to assess whether the instrument used was appropriate for the intended use. The checklist provides simplified criteria that can be used to assess developmental properties and usefulness of a variety of PRO measures by end users with a wide range of expertise in measurement theory, psychometrics, or survey development. Our intent was not to replace the currently available comprehensive evaluative consensus guidelines. Instead, we propose that these criteria serve as a distilled and simplified version of characteristics that constitute an adequately developed PRO measure. Psychometricians, statisticians, measurement theory experts, econometricians, and clinicians have iteratively developed and discussed these properties over decades, in a literature that encompasses an array of disciplines. Refinements and evolution of these techniques continue. However, the general fundamentals remain the bedrock on which these innovations build. Our criteria attempt to summarize these foundational concepts into a user-friendly checklist that will help end users with a variety of backgrounds to identify the strengths and weaknesses of available PRO measures for their particular application.

Additional file

Additional file 1: Written description of patient-reported outcome measure concepts included in the checklist. (DOCX 16.0 MB)

Acknowledgements

Funding was provided by National Institute of Health—National Institute for Deafness and Communication Disorders K23DC013559 (DOF) and L30DC012687 (DOF).

Authors' contributions

DOF was involved in the study conception, design, and development of the checklist and the collection and analysis of data and drafted the manuscript. MM participated in the study conception and study design and helped in drafting the manuscript. MN coordinated contacts for the cognitive interviews and helped in compilation of the data and was involved in drafting. DP assisted in the study conception and design, analysis, and drafting of the manuscript, and IF was central and worked in conjunction with DOF in the study conception and design, development of the checklist, analysis of data, and drafting of the manuscript. All authors approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Otolaryngology, Vanderbilt University Medical Center, Medical Center East, Suite 7302, 1215, 21st Avenue South, Nashville, TN 37212, USA. [2]Center for Surgical Quality and Outcomes Research, Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville 37232, TN, USA. [3]Vanderbilt Evidence-based Practice Center, Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville 37232, TN, USA. [4]Department of Health Policy, Vanderbilt University Medical Center, Nashville 37232, TN, USA. [5]Center of Population Science, Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville 37232, TN, USA. [6]Departments of Urological Surgery and Medicine, Vanderbilt University Medical Center, Nashville 37232, TN, USA. [7]Geriatric Research Education and Clinical Center, Veterans Administration Tennessee Valley Healthcare System, Nashville, USA. [8]Departments of Surgery and Biostatistics, Vanderbilt University Medical Center, Nashville 37232, TN, USA.

Received: 19 December 2015 Accepted: 20 June 2016
Published online: 02 August 2016

References

1. Rosser R. A history of the development of health indicators. In: Teeling-Smith G, editor. Measure the social benefits of medicine. London: Office of Health Economics; 1983. p. 50–63.
2. Goldsmith SB. The status of health status indicators. Health Serv Rep. 1972; 87(3):212–20.
3. Goldsmith SB. A reevaluation of health status indicators. Health Serv Rep. 1973;88(10):937–41.
4. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 2nd ed. New York: Oxford University Press; 1996. xix, 523 pp.
5. Patrick DL, Burke LB, Powers JH, Scott JA, Rock EP, Dawisha S, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health. 2007;10 Suppl 2:S125–37.
6. Guidance for Industry. Patient-reported outcome measures: use in medical product development to support labeling claims. Rockville, MD: 2009
7. Patient-Centered Outcomes Research Institute. Rationale: working definition of patient-centered outcomes research 2014. Available from: http://www.pcori.org/research-results/patient-centered-outcomes-research. Accessed 25 July 2016.
8. 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 Int J Qual Life Asp Treat Care Rehab. 2013;22(8):1889–905.
9. Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients? Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2007;16(7):1097–105.
10. Feeny DH, Eckstrom E, Whitlock EP, Perdue LA. A primer for systematic reviewers on the measurement of functional status and health-related quality of life in older adults (prepared by the Kaiser Permanente Affiliates Evidence-based Practice Center under contract 290-2007-10057-I.) ARHQ Publication No. 13-EHC128-EF. Rockville: Agency for Healthcare Research and Quality; 2013
11. 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(7):737–45.
12. Patrick DL, Erickson P. Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford University Press; 1993. xxv, 478 pp.
13. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med. 1993;118(8):622–9.
14. Patrick DL, Bergner M. Measurement of health status in the 1990s. Annu Rev Public Health. 1990;11:165–83.
15. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. JAMA. 1994;272(8):619–26.
16. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol. 2006;6:2.
17. Weldring T, Smith SM. Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights. 2013;6:61–8.
18. Guyatt G, Montori V, Devereaux PJ, Schunemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. ACP J Club. 2004;140(1):A11–2.
19. Garrow AP, Khan N, Tyson S, Vestbo J, Singh D, Yorke J. The development and first validation of the Manchester Early Morning Symptoms Index (MEMSI) for patients with COPD. Thorax. 2015;70(8):757–63.
20. Hutchings HA, Cheung WY, Russell IT, Durai D, Alrubaiy L, Williams JG. Psychometric development of the Gastrointestinal Symptom Rating Questionnaire (GSRQ) demonstrated good validity. J Clin Epidemiol. 2015; 68(10):1176–83.
21. Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. BMJ. 1993; 306(6890):1437–40.
22. Cuervo J, Castejon N, Khalaf KM, Waweru C, Globe D, Patrick DL. Development of the Incontinence Utility Index: estimating population-based utilities associated with urinary problems from the Incontinence Quality of Life Questionnaire and Neurogenic Module. Health Qual Life Outcomes. 2014;12:147.
23. Andrae DA, Patrick DL, Drossman DA, Covington PS. Evaluation of the Irritable Bowel Syndrome Quality of Life (IBS-QOL) questionnaire in diarrheal-predominant irritable bowel syndrome patients. Health Qual Life Outcomes. 2013;11:208.

24. Wang Q, Furlong W, Feeny D, Torrance G, Barr R. How robust is the Health Utilities Index Mark 2 utility function? Med Decis Making. 2002;22(4):350–8.
25. Kopec JA, Schultz SE, Goel V, Ivan WJ. Can the health utilities index measure change? Med Care. 2001;39(6):562–74.
26. Pouwer F, Snoek FJ, van der Ploeg HM, Ader HJ, Heine RJ. The well-being questionnaire: evidence for a three-factor structure with 12 items (W-BQ12). Psychol Med. 2000;30(2):455–62.
27. Monk M. Blood pressure awareness and psychological well-being in the health and nutrition examination survey. Clin Invest Med. 1981;4(3–4):183–9.
28. Sherbourne CD, Stewart AL. The MOS social support survey. Soc Sci Med. 1991;32(6):705–14.
29. Riehm KE, Kwakkenbos L, Carrier ME, Bartlett SJ, Malcarne VL, Mouthon L, et al. Validation of the Self-Efficacy for Managing Chronic Disease (SEMCD) scale: a Scleroderma Patient-centered Intervention Network (SPIN) cohort study. Arthritis care & research. 2016;8(68):1195–200.
30. Pleil AM, Coyne KS, Reese PR, Jumadilova Z, Rovner ES, Kelleher CJ. The validation of patient-rated global assessments of treatment benefit, satisfaction, and willingness to continue—the BSW. Value Health. 2005;8 Suppl 1:S25–34.
31. Wegener ST, Castillo RC, Heins SE, Bradford AN, Newell MZ, Pollak AN, et al. The development and validation of the readiness to engage in self-management after acute traumatic injury questionnaire. Rehabil Psychol. 2014;59(2):203–10.
32. Health USDo, Human Services FDACfDE, Research, Health USDo, Human Services FDACfBE, Research, et al. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. 2006;4:79.
33. Johnston BC, Ebrahim S, Carrasco-Labra A, Furukawa TA, Patrick DL, Crawford MW, et al. Minimally important difference estimates and methods: a protocol. BMJ Open. 2015;5(10), e007953.
34. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2012;21(4):651–7.
35. Scoggins JF, Patrick DL. The use of patient-reported outcomes instruments in registered clinical trials: evidence from ClinicalTrials.gov. Contemp Clin Trials. 2009;30(4):289–92.
36. Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract. 2000;6(1):39–49.
37. Zheng H, Li W, Harrold L, Ayers DC, Franklin PD. Web-based comparative patient-reported outcome feedback to support quality improvement and comparative effectiveness research in total joint replacement. EGEMS (Wash DC). 2014;2(1):1130.
38. Black N, Burke L, Forrest CB, Ravens Sieberer UH, Ahmed S, Valderas JM, et al. Patient-reported outcomes: pathways to better health, better services, and better societies. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2016;25(5):1103–12. doi:10.1007/s11136-015-1168-3. Epub 2015 Nov 13.
39. Basch E, Spertus J, Dudley RA, Wu A, Chuahan C, Cohen P, et al. Methods for developing patient-reported outcome-based performance measures (PRO-PMs). Value Health. 2015;18(4):493–504.
40. Anastasi A. Psychological testing. New York: Macmillian Publishing Company; 1988.
41. Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med. 1983;99(6):843–8.
42. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.
43. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2002;11(3):193–205.
44. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2010;19(4):539–49.
45. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2009;18(3):313–33.
46. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010;10:22.
47. Regnault A, Hamel JF, Patrick DL. Pooling of cross-cultural PRO data in multinational clinical trials: how much can poor measurement affect statistical power? Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2015;24(2):273–7.
48. Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol. 2010;10:82.
49. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther. 1996;18(5):979–92.
50. American Psychological Association AERA, National Council on Measurement Used in Education. Technical recommendations for psychological tests and diagnostic techniques. Psychol Bull. 1954;51(2, pt 2):1–38.
51. American Psychological Association AERA. Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association; 1999.
52. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford: New York: Oxford University Press; 2008. xvii, 431 pp.
53. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. Oxford; New York: Oxford University Press; 2006. xvi, 748 pp.
54. Spilker B. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996. xlv, 1259 pp.
55. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014;36(5):648–62.
56. Allen MJ, Yen WM. Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company; 1979.
57. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J Clin Epidemiol. 1992;45(11):1201–18.
58. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273(1):59–65.
59. Dahl TH. International classification of functioning, disability and health: an introduction and discussion of its potential impact on rehabilitation services and research. J Rehabil Med. 2002;34(5):201–4.
60. World Health Organization. International classification of functioning, disability, and health. Geneva: World Health Organization; 2001.
61. Magasi S, Ryan G, Revicki D, Lenderking W, Hays RD, Brod M, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 2012;21(5):739–46.
62. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. Value Health. 2011; 14(8):967–77.
63. Helmstadter GC. Principles of psychological measurement. New York: Appleton; 1964. xx, 248 pp.
64. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care. 2000;38(9 Suppl):II84–90.
65. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res Int J Qual Life Asp Treat Care Rehab. 1992;1(1):73–5.
66. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102–9.
67. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171–8.
68. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15.
69. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol. 1994;47(1):81–7.
70. Idler EL, Kasl SV, Lemke JH. Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982–1986. Am J Epidemiol. 1990;131(1):91–103.

71. Kroenke K, Monahan PO, Kean J. Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. J Clin Epidemiol. 2015;68(9):1085–92.
72. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting G. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371–83.
73. Schunemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). COPD. 2005;2(1):81–9.
74. Ware Jr JE, Keller SD, Hatoum HT, Kong SX. The SF-36 Arthritis-Specific Health Index (ASHI): I. Development and cross-validation of scoring algorithms. Med Care. 1999;37(5 Suppl):MS40–50.
75. Schunemann HJ, Guyatt GH. Commentary—goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005;40(2):593–7.
76. Brozek JL, Guyatt GH, Schunemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes. 2006;4:69.
77. Meyer KB, Clayton KA. Measurement and analysis of patient-reported outcomes. Methods Mol Biol. 2009;473:155–69.
78. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 2013;309(8):814–22.
79. Chenok K, Teleki S, SooHoo NF, Huddleston 3rd J, Bozic KJ. Collecting patient-reported outcomes: lessons from the California Joint Replacement Registry. EGEMS (Wash DC). 2015;3(1):1196.
80. Jahagirdar D, Kroll T, Ritchie K, Wyke S. Patient-reported outcome measures for chronic obstructive pulmonary disease: the exclusion of people with low literacy skills and learning disabilities. The Patient. 2013;6(1):11–21.
81. Jacobson B, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS. The Voice Handicap Index (VHI): development and validation. Am J Speech-Language Pathology. 1997;6:66–70.
82. Rosen CA, Lee AS, Osborne J, Zullo T, Murry T. Development and validation of the voice handicap index-10. Laryngoscope. 2004;114(9):1549–56.
83. Ma EP, Yiu EM. Voice activity and participation profile: assessing the impact of voice disorders on daily activities. Journal of speech, language, and hearing research : JSLHR. 2001;44(3):511–24.
84. Epstein R, Hirani SP, Stygall J, Newman SP. How do individuals cope with voice disorders? Introducing the Voice Disability Coping Questionnaire. J Voice. 2009;23(2):209–17.
85. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). J Voice. 1999;13(4):557–69.
86. Urbach DR, Tomlinson GA, Harnish JL, Martino R, Diamant NE. A measure of disease-specific health-related quality of life for achalasia. Am J Gastroenterol. 2005;100(8):1668–76.
87. Ma EP, Yiu EM. Scaling voice activity limitation and participation restriction in dysphonic individuals. Folia Phoniatr Logop. 2007;59(2):74–82.
88. Feinstein AR. Benefits and obstacles for development of health status assessment measures in clinical settings. Med Care. 1992;30(5 Suppl):MS50–6.
89. Ahmed S, Berzon RA, Revicki DA, Lenderking WR, Moinpour CM, Basch E, et al. The use of patient-reported outcomes (PRO) within comparative effectiveness research: implications for clinical practice and health care policy. Med Care. 2012;50(12):1060–70.
90. Newton PE, Shaw SD. Standards for talking and thinking about validity. Psychol Methods. 2013;18(3):301–19.