JSLHR

Review Article

Voice-Related Patient-Reported Outcome Measures: A Systematic Review of Instrument Development and Validation

David O. Francis,[a,b,c] James J. Daniero,[d] Kristen L. Hovis,[e] Nila Sathe,[c,f,g] Barbara Jacobson,[h] David F. Penson,[b,c,g,i,j] Irene D. Feurer,[b,k] and Melissa L. McPheeters[c,g]

Purpose: The purpose of this study was to perform a comprehensive systematic review of the literature on voice-related patient-reported outcome (PRO) measures in adults and to evaluate each instrument for the presence of important measurement properties.

Method: MEDLINE, the Cumulative Index of Nursing and Allied Health Literature, and the Health and Psychosocial Instrument databases were searched using relevant vocabulary terms and key terms related to PRO measures and voice. Inclusion and exclusion criteria were developed in consultation with an expert panel. Three independent investigators assessed study methodology using criteria developed a priori. Measurement properties were examined and entered into evidence tables.

Results: A total of 3,744 studies assessing voice-related constructs were identified. This list was narrowed to 32 PRO measures on the basis of predetermined inclusion and exclusion criteria. Questionnaire measurement properties varied widely. Important thematic deficiencies were apparent: (a) lack of patient involvement in the item development process, (b) lack of robust construct validity, and (c) lack of clear interpretability and scaling.

Conclusions: PRO measures are a principal means of evaluating treatment effectiveness in voice-related conditions. Despite their prominence, available PRO measures have disparate methodological rigor. Care must be taken to understand the psychometric and measurement properties and the applicability of PRO measures before advocating for their use in clinical or research applications.

Voice disorders have an estimated point prevalence of 20 million (0.98%) in the United States (Cohen, Kim, Roy, Asche, & Courey, 2012b; Roy, Merrill, Gray, & Smith, 2005). Annual direct costs exceed $5 billion even before accounting for productivity losses due to absenteeism and presenteeism (Cohen, Kim, Roy, Asche, & Courey, 2012a; Dew, Keefe, & Small, 2005). Quality of life consequences for voice disorders have a magnitude similar to that of chronic sinusitis, sciatica, and angina pectoris (Benninger, Ahuja, Gardner, & Grywalski, 1998). A need exists to improve care, but this requires the ability to quantify a given voice disorder's effect on the patient.

Several categories of voice measurement are used in clinical practice, such as the Consensus Auditory Perceptual Evaluation of Voice (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009); the Grade, Roughness, Breathiness, Asthenia, Strain scale (Hirano, 1981); laryngoscopy; and patient-reported outcome (PRO) measures. Of central importance are PRO measures, defined as "any report of 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" (Guyatt

[a]Vanderbilt Voice Center, Department of Otolaryngology, Bill Wilkerson Center, Vanderbilt University Medical Center, Nashville, TN
[b]Center for Surgical Quality and Outcomes Research, Nashville, TN
[c]Vanderbilt Evidence-Based Practice Center, Nashville, TN
[d]Center for Voice and Swallowing, University of Virginia, Charlottesville
[e]Vanderbilt University School of Medicine, Nashville, TN
[f]Vanderbilt University Medical Center, Nashville, TN
[g]Department of Health Policy, Vanderbilt University Medical Center, Nashville, TN
[h]Department of Hearing and Speech Sciences, Bill Wilkerson Center, Vanderbilt University Medical Center, Nashville, TN
[i]Departments of Urology and Medicine, Vanderbilt University Medical Center, Nashville, TN
[j]Geriatric Research and Education Coordination Center, Veteran's Administration Tennessee Valley Health System, Geriatric Research and Education Coordination Center, Nashville, TN
[k]Departments of Surgery and Biostatistics, Vanderbilt University Medical Center, Nashville, TN

Correspondence to David O. Francis: david.o.francis@vanderbilt.edu

Editor: Julie Liss
Associate Editor: Nelson Roy

Received January 20, 2016
Revision received April 28, 2016
Accepted July 20, 2016
DOI: 10.1044/2016_JSLHR-S-16-0022

Disclosure: The authors have declared that no competing interests existed at the time of publication.

& Schunemann, 2007; Reeve et al., 2013; Snyder, Jensen, Segal, & Wu, 2013), which provide a method of systematically capturing patient perspective and experience. Including the patient experience related to treatment benefit and harm is now obligatory in the United States' preapproval regulatory setting. In fact, the U.S. Food and Drug Administration (FDA) recommends the use of PRO measures as part of the approval process for pharmacological products and devices (Ahmed et al., 2012). PRO measures have also been highlighted in the National Institutes of Health Roadmap, which identifies priority areas that have the greatest potential to drive progress in biomedical research (Patient-Centered Outcomes Research Institute, 2014).

PRO measures are increasingly used to better understand the perspectives of, and to measure concepts that matter to, the patient (Patrick et al., 2007). A National Institutes of Health/FDA working group identified three patient-centered outcome categories—feeling, function, and survival—as primary outcomes to be focused on and incorporated into all clinical trials proposing novel interventions, devices, or pharmaceuticals that aim for FDA approval (Patrick et al., 2007). Several conceptual models for PRO measure development exist, including the International Classification of Functioning, Disability and Health promoted by the World Health Organization (2001). The International Classification of Functioning, Disability and Health paradigm is considered the international standard for conceptualizing the measurement of health and disability. Its basic tenets have formed the rubric for PRO measures in a wide variety of health topics, including voice and voice disorders (Hogikyan & Sethuraman, 1999; Jacobson et al., 1997). It is important to note that although it does provide a conceptual framework, it does not provide guidance on how to develop a PRO measure.

There is also a misconception that these instruments are designed only to measure health-related quality of life. In reality, they can be designed as symptom indices (Belafsky, Postma, & Koufman, 2002; Garrow et al., 2015); measure general (McHorney, Ware, & Raczek, 1993; Ware & Sherbourne, 1992) or condition-specific (Andrae, Patrick, Drossman, & Covington, 2013; Hutchings et al., 2015) health-related quality of life, utility (Feeny, Furlong, Boyle, & Torrance, 1995; Torrance, Furlong, Feeny, & Boyle, 1995), well-being (Monk, 1981; Pouwer, Snoek, van der Ploeg, Ader, & Heine, 2000), or social health (Sherbourne & Stewart, 1991); or can focus on latent structures such as self-efficacy (Riehm et al., 2016) and willingness to change (Pleil et al., 2005; Wegener et al., 2014). Increased focus on patient-centered outcomes research has resulted in a proliferation of PRO measures with variable psychometric rigor (Johnston et al., 2015).

Methodological experts in measurement theory and survey design have disseminated several consensus statements to guide appropriate development and implementation of these measures (Aaronson et al., 2002; Feeny, Eckstrom, Whitlock, & Perdue, 2013; Mokkink et al., 2010; Patrick et al., 2007; Reeve et al., 2013; Terwee et al., 2007). Use of a poorly developed PRO measures or those designed for a purpose that differs from its use can have significant implications and lead to distorted, inaccurate, or equivocal findings (Mokkink et al., 2009, 2010; Regnault, Hamel, & Patrick, 2015). Measures should be chosen on the basis of relevance and their track record in the context of the proposed study. 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.

To date, no criterion objective test supersedes the importance of the patient's perspective in the evaluation of voice disorders. In fact, objective metrics used in clinical practice have correlated poorly with subjective patient-reported improvement and current PRO measures (Cheng & Woo, 2010; Hsiung, Pai, & Wang, 2002; Wheeler, Collins, & Sapienza, 2006; Woisard, Bodin, Yardeni, & Puech, 2007). Although a complete voice evaluation requires some level of objective assessment, results should be contextualized within each patient's perspective and expectations (Roy et al., 2013). For instance, some people have known only a rough voice. This is their "normal" despite acoustic, cepstral, audio-perceptual, and visual-perceptual assessments indicating a disordered voice.

Practitioners recognize that if patients are satisfied with their voices and no negative health consequence exists (e.g., malignancy), then further intervention (e.g., surgery, voice therapy) is difficult to justify even if the practitioner rates the voice as disordered. As such, PRO measures are arguably the primary tools for systematically assessing both the individual's perspective and population-level burden of voice-related disease. The Cochrane Collaboration routinely uses PRO measures as primary outcome assessments in its systematic reviews of associated topics (Hopkins, Yousaf, & Pedersen, 2006; Ruotsalainen, Sellman, Lehto, Jauhiainen, & Verbeek, 2007a, 2007b).

Recognition of the importance of the patient-centered approach has led to the proliferation of PRO measures for a variety of voice-related constructs. Methodological inconsistency exists because of the complexity of proper PRO measure development. The use of questionnaires that do not adhere to meticulous measure-instrument construction principles—for example, psychometrics (Anastasi, 1988) or clinimetrics (Feinstein, 1983)—can yield spurious data and incorrect conclusions (Penson, Litwin, & Aaronson, 2003). To date, three systematic reviews of voice-related PRO measures have been published. One was performed as part of a larger 2002 Agency for Healthcare Research and Quality systematic review, at which time (1996–2000) few voice-related PRO measures existed (Biddle, Watson, Hooper, Lohr, & Sutton, 2002). A subsequent review found that all five voice quality of life instruments that it identified were incomplete in their psychometric development (Franic, Bramlett, & Bothe, 2005). Another found content validity inadequacies in all nine of the questionnaires it identified (Branski et al., 2010). These reviews had different objectives and assessed few measures currently available in the literature. Differences in number of identified studies likely relates to earlier publication dates, a specific focus on quality of life instruments (perhaps resulting in exclusion of PRO measures with different constructs; e.g., symptom severity, self-efficacy), and variable literature search strategies.

The intent of the present systematic review is to assess the measurement characteristics of all currently available adult voice-related PRO measures in order to identify their strengths, weaknesses, and applicability. It aims to evaluate each instrument's measurement properties, including conceptual model, content validity, reliability, construct validity, responsiveness to change, scoring and interpretation, and respondent burden and presentation. These important parameters have significant ramifications for the applicability of PRO measures. The ultimate goal of the present study is to fill these voids and to provide guidance in terms of how to evaluate a PRO measure and to aid in selection of an appropriate instrument for a specific application.

Methods

Search Strategy

MEDLINE via the PubMed interface, the Cumulative Index of Nursing and Allied Health Literature, and the Health and Psychosocial Instrument database were searched using relevant vocabulary terms and key terms related to PRO measures and voice (see Appendixes A–C). No restrictions on publication date were used. The initial literature search was conducted in November 2014 and was updated in April 2015. Reference lists of the included articles and recent reviews related to measurement of voice were hand searched to identify additional relevant articles.

Study Selection

Inclusion and exclusion criteria were developed in consultation with an expert panel that included a statistician with expertise in measurement theory (the seventh author), systematic review methodologists (the fourth and eighth authors), and researchers and clinicians who treat and study voice and voice disorders (the first, second, and fifth authors). Abstracts for all studies identified in the literature search were independently reviewed by three investigators (the first, second, and third authors), and those meeting predetermined abstract screening criteria (see Table 1) were advanced to full-text review. Measures focused on singing voice and pediatric voice were excluded. Articles lacking adequate information in their title or abstract to determine eligibility were also included in the full-text review phase. Three independent reviewers performed full-text review of articles to determine eligibility for data extraction. Disagreements were resolved through discussion or adjudication by a senior investigator (the seventh author). When necessary, article authors were personally contacted for further information.

Table 1. Screening criteria for abstract review.

Original research (includes systematic reviews and meta-analysis but not narrative reviews)?
Research is on human subjects?
Study addresses voice problems or hoarseness?
Study addresses a patient-reported outcome, instrument, questionnaire, or survey?
Study addresses development, validity testing, and/or reliability testing of a patient-reported outcome measure, instrument, survey, or questionnaire?
Study performed in adult population (≥18 years of age)?

Data Extraction

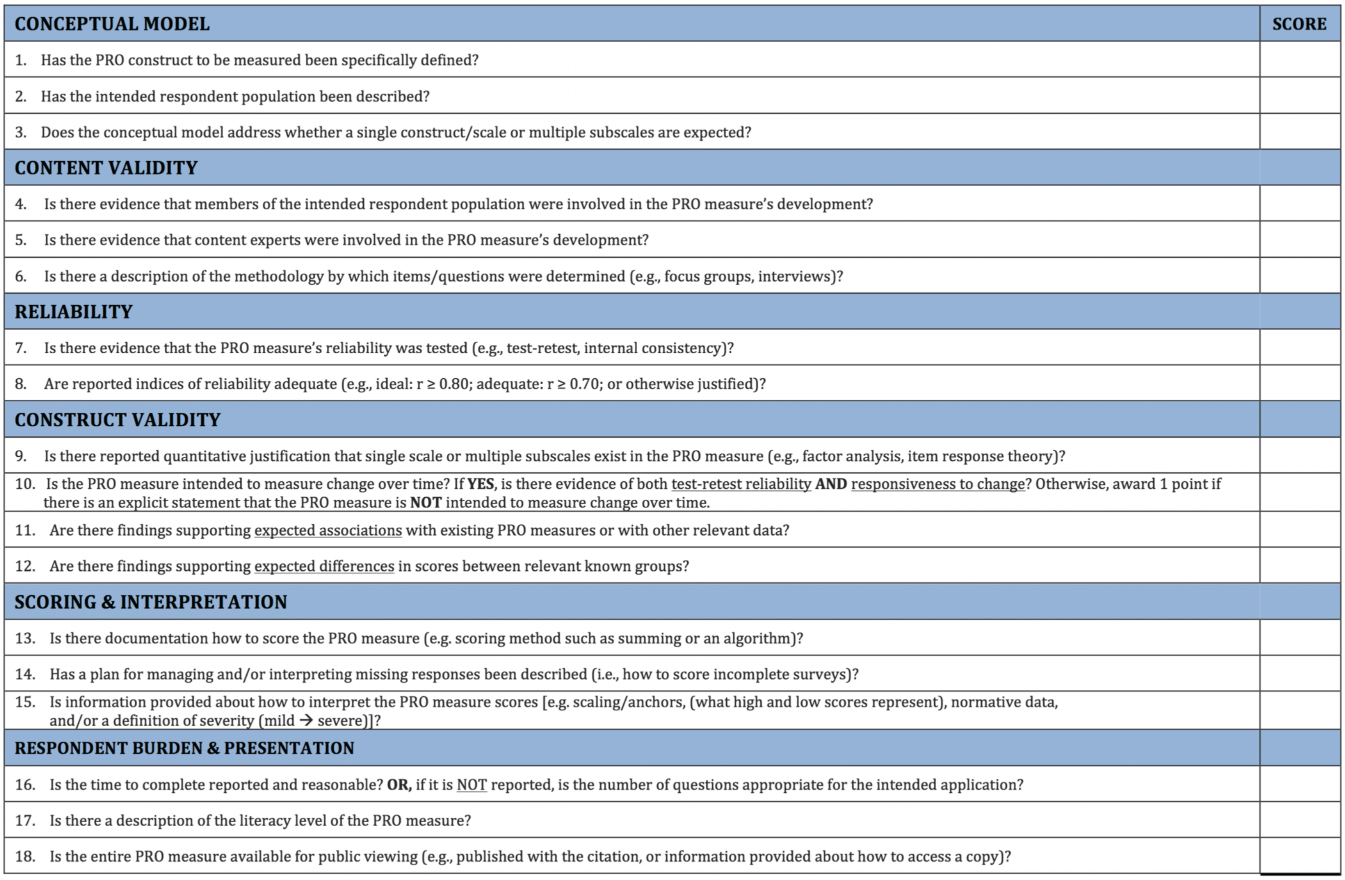
One reviewer extracted all relevant data from studies meeting criteria at the full-text review phase. A second reviewer independently verified data accuracy. Components of PRO measure development were critically examined and entered into evidence tables. These included PRO measure name and acronym, authors, year published, objective and intended construct, setting of development (e.g., tertiary care, community) and country, population targeted and involved in instrument development, type of scale used (e.g., Likert, visual analog scale), number of items or questions, and, when present, what subscales or domains they were designed to specifically measure.

PRO Measure Assessment

Three investigators independently assessed each study's methodology using a criteria checklist developed a priori (see Figure 1; Francis, McPheeters, Noud, Penson, & Feurer, 2016). In brief, the checklist used was designed as a tool to help systematic reviewers identify components considered important in the development of questionnaires. The checklist helps users evaluate a prospective PRO measure's conceptual model, content validity, reliability, construct validity, responsiveness to change, scoring and interpretation, and respondent burden and presentation. A glossary of important measurement properties is shown in Table 2. This tool is not meant to yield a total score, as that implies equal weighting of included items. Instead, it is intended as a guide to identify whether important measurement properties are present in current PRO measures. Each item is scored in a dichotomous manner (i.e., presence or absence of a component) and does not attempt to grade the quality of particular parameters.

For this study, each reviewer was trained and calibrated on appropriate application of the checklist using a methodology described separately (Francis et al., 2016). Each reviewer reviewed six voice and swallowing PRO measures with variable psychometric construction approaches and different measurement properties without instruction. Their scoring was compared to that of individuals with extensive experience in instrument development and psychometrics. If agreement was insufficient on the first pass, they received directed education and repeated the scoring process. Reviewers for this study needed one round of instruction, after which they demonstrated near-perfect agreement with experts. Once determined to be competent, they were then independently tasked with evaluating all identified

Figure 1. Checklist of key characteristics to consider when evaluating a patient-reported outcome (PRO) measure. Indicate in the score column whether the information provided in the citation/source document meets each of the criteria (0 = *criterion not met*, 1 = *criterion met*).



CONCEPTUAL MODEL	**SCORE**
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. 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.	
11. Are there findings supporting expected associations with existing PRO measures or with other relevant data?	
12. Are there findings supporting expected differences in scores between relevant known groups?	
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)?	

voice-related PRO measures. Upon completion, reviewers met to discuss and come to consensus on scoring discrepancies. Initial agreement was greater than 75% among reviewers for all but three parameters: justification of subscales (72%), longitudinal validity (75%), and description of item development (75%; see Table 3). All discrepancies were discussed, and articles in question were reviewed together until consensus was achieved. A senior psychometrician (the seventh author) adjudicated the few remaining discrepancies.

Data Synthesis

Data from unique PRO measures demonstrated wide heterogeneity in constructs, methodology, and intended purpose. Thus, data were not appropriate for aggregation or meta-analysis. Instead, individual PRO characteristics were summarized independently with respect to instrument construction and psychometric rigor.

Results

Figure 2 is the Preferred Reporting Items for Systemic Reviews and Meta-Analyses (PRISMA) diagram describing the study flow and inclusions. The most common reasons for excluding articles were that they lacked relevance, did not describe de novo development or validation of an existing PRO measure, or involved a primarily pediatric population. A total of 34 studies were identified that provided initial development process data on 32 voice-related PRO measures (see Tables 4 and 5).

Publication year ranged from 1984 (Linear Analog Scale Assessment of Voice Quality [LASA-VQ]; Llewellyn-Thomas et al., 1984) to 2015 (Vocal Fatigue Index [VFI]; Nanjundeswaran, Jacobson, Gartner-Schmidt, & Verdolini Abbott, 2015), with increasingly more instruments being introduced over time (see Figure 3). In order of frequency, PRO measures were developed in the United States (13), Great Britain (7), the Netherlands (3), Brazil (2), Italy (2), Canada (1), Hong Kong (1), Finland (1), India (1), and South Korea (1; see Table 5). Development of each instrument occurred at an academic center. Sample size used in the instrument development process varied from nine to 1,310 subjects (see Table 5). One measure was a subscale within a broader instrument: Scleroderma Logopedic Scale (SLS-Voice; Vitali et al., 2010). Ten studies did not report the age and/or gender distribution of respondents (cases and controls, when applicable) used in each step of PRO measure development: Voice-Related Quality of Life (V-RQOL;

Table 2. Glossary of measurement properties of patient-reported outcome (PRO) measures.

Domain	Explanation
Conceptual model	A conceptual model provides a rationale for and description of the concepts and target population that a measure is intended to assess.
Content validity	Content validity refers to evidence that a PRO measure's domain(s) are appropriate for its intended use. Items and conceptual domains should be relevant to the target population's concerns. The PRO measure's development should include direct input from patients and from content experts. There should be a clear description of the process by which included questions were derived.
Reliability	Reliability is the degree to which scores are free from random (measurement) error. Internal consistency reliability, the degree to which segments of a test (e.g., individual items) are associated with one another, reflects precision at a single time point. Test–retest reliability refers to the reproducibility of scores over two administrations, typically in close temporal proximity, among respondents who are assumed not to have changed on the relevant domains. Cited minimum levels for reliability coefficients traditionally are .70 for group-level comparisons and .90 to .95 for individual comparisons. Reliability estimates lower than these conventions should be justified in the context of the proposed PRO measure's intended application.
Construct validity	Construct validity refers to whether a test measures intended theoretic constructs or traits and directly affects the appropriateness of the measurement-based inferences. Several different forms exist and are outlined below. Empirical demonstration of dimensionality (e.g., factor analysis) provides evidence of whether a single scale or multiple subscales exist in the PRO measure. Responsiveness to change (longitudinal validity) is the extent to which a PRO measure detects meaningful change over time when it is known to have occurred. It is predicated on demonstration of both test–retest-reliability (stability when no change is expected) and clinically meaningful change when it is expected. Convergent validity is the degree to which a PRO measure's scores correlate with other instruments that measure the same construct or with related clinical indicators (e.g., diagnostic test). A priori hypotheses about expected associations between a PRO measure and similar or dissimilar measures should be documented. Known-groups validity is the degree to which a PRO measure is able to differentiate between groups that empiric evidence has shown to be different (e.g., cases and controls).
Interpretability and scoring	Interpretability is the degree to which the meaning of the scores can be easily understood. Scoring refers to the "rules" for computing total scores or scales, if relevant. A description of how to score the measure (e.g., summation, algorithm) should be provided. Missing responses are a common occurrence in both clinical and research settings and can affect an end user's ability to interpret results. A prespecified plan for managing missing responses can mitigate the risk of bias resulting from the necessity to exclude cases with missing data. Scaling is the process of distributing the full range of respondents' possible scores with respect to the measured attribute. A relative score then represents a subject's location in relation to others on a common scale. It allows cross-sectional and longitudinal quantification of the magnitude of the attribute that is reported and its change over time. Both cross-sectional and longitudinal changes in scores need to be contextualized to allow interpretation of their meaning. Ideally, scaling should be based on an understanding of what represents a clinically important or patient-important change in the construct being measured.
Burden and presentation	Burden refers to the time, effort, or other demands placed on respondents or those administering the instrument. This includes number and complexity of items. The literacy level needed to understand and complete the measure is another important aspect of burden. Although most experts recommend literacy be at the sixth-grade reading level or lower, this criterion should be contextualized to the intended target population. Presentation refers to a questionnaire's appearance in light of its intended mode of administration. It is important that prospective users be able to preview a measure in its entirety (e.g., items and response options) to ensure its appropriateness for the intended application.

Hogikyan & Sethuraman, 1999), Voice Activity and Participation Profile (VAPP; Ma & Yiu, 2007), Glottal Function Index (GFI; Bach, Belafsky, Wasylik, Postma, & Koufman, 2005), Voice Handicap Index-10 (VHI-10; Arffa, Krishna, Gartner-Schmidt, & Rosen, 2012; Rosen, Lee, Osborne, Zullo, & Murry, 2004), Voice Self-Efficacy Questionnaire (VSEQ; Gillespie & Abbott, 2011), VFI, Screening Index for Voice Disorders (SIVD; Ghirardi, Ferreira, Giannini, & Latorre Mdo, 2013), Self-Efficacy in Spasmodic Dysphonia (SE-SD; Hu et al., 2013), SLS-Voice, and Voice Disorder Outcome Profile (Voice-DOP; Konnai, Jayaram, & Scherer, 2010). Distribution of pathology among respondents differed by PRO measure. Among voice-related PRO measures, only one of 32 used item response theory psychometric techniques (Communicative Participation Item Bank [CPIB]; Baylor et al., 2013, 2014; Baylor, Yorkston, Eadie, Miller, & Amtmann, 2009; Eadie et al., 2014), whereas the remainder applied clinimetric (Feinstein, 1983) or classical test theory (Anastasi, 1988) methodology.

Constructs Measured

The constructs measured were heterogeneous (see Table 4) and included the following:

Table 3. Initial rater agreement for patient-reported outcome measure assessment domain and criterion.

Domain	Criterion	Initial agreement (%)
Conceptual model	Construct defined	97
	Target population defined	100
	Expected subscales defined	81
Content validity	Patients devised items	78
	Content experts involved	100
	Description of item development	75
Reliability	Reliability tested	94
	Coefficients adequate	78
Construct validity	Justification of subscales	72
	Convergent validity	81
	Known-group validity	84
Responsiveness	Longitudinal validity	75
Interpretation and scoring	Plan for scoring measure	84
	Plan for missing data	84
	Scaling described	78
Burden and presentation	Length reasonable	94
	Literacy level	97
	Items viewable	91

- Coping: Voice Disability Coping Questionnaire (VDCQ; Epstein, Hirani, Stygall, & Newman, 2009)
- Quality of life: Voice Outcome Survey (VOS; Gliklich, Glovsky, & Montgomery, 1999), V-RQOL, Voice-DOP, Evaluating Voice Disability–Quality of Life Questionnaire (EVD-QOL; Smith et al., 1996), 3-Item Outcome Scale (3-IOS; Speyer, Wieneke, & Dejonckere, 2004)
- Handicap: Voice Handicap Index (VHI; Jacobson et al., 1997; Rosen, Murry, Zinn, Zullo, & Sonbolian, 2000), VHI-10, CPIB
- Vocal performance: Vocal Performance Questionnaire (VPQ; Carding & Horsley, 1992; Carding, Horsley, & Docherty, 1998)
- Vocal impairment: Self-Ratings of Vocal Performance (SRVP; Verdonck-de Leeuw et al., 1999)
- Vocal fatigue: Self-Evaluation of Voice as Treatment Outcome Measure (SEVTOM; Laukkanen, Leppanen, & Ilomaki, 2009), VFI, Vocal Fatigue Handicap Questionnaire (VFHQ; Paolillo & Pantaleo, 2015)
- Voice quality: LASA-VQ, Thyroidectomy-Related Voice Questionnaire (TVQ; Nam et al., 2012)
- Self-efficacy: SE-SD, VSEQ
- Work productivity: Work Productivity Activity Impairment Questionnaire–Specific Health Problem–Voice (WPAI-SHP; Isetti & Meyer, 2014), Stanford Presenteeism Scale 6 (SPS-6; Isetti & Meyer, 2014),

Figure 2. Number and acronyms of new voice-related patient-reported outcome measures over time.

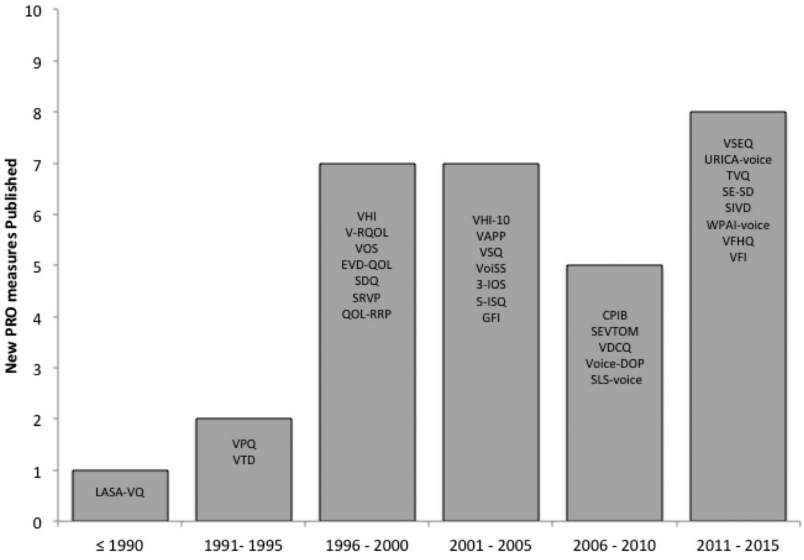


Table 4. Measurement aims, target populations, and item characteristics of voice-related patient-reported outcome measures.

Patient-reported outcome measure	Year	Measurement aim	Target population	Language[a]	Number of items/ subscales	Response options	Subscales
Linear Analog Scale of Assessment Voice Quality (LASA-VQ)	1984	To determine whether patients' self-assessment of voice gives reliable data and whether this method of assessment would be sensitive to clinical change occurring during the course of radiation therapy	Individuals with laryngeal cancer	English	16 items in 2 subscales	Visual analog scales	Vocal symptoms Functional abilities
Vocal Performance Questionnaire (VPQ)	1992	To investigate the effectiveness of speech therapy in the treatment of nonorganic dysphonia	Individuals with nonorganic dysphonia	English	12 items	5-point Likert scale (a–e)	
Vocal Tract Discomfort (VTD)	1993	(a) To determine incidence of vocal tract discomfort in a group of patients with hyperfunctional dysphonia; (b) to assess qualitative differences in the discomfort experienced; (c) to correlate between discomfort and vocal fold mucosal damage; (d) to assess discomfort resolution time during treatment	Individuals with hyperfunctional voice disorders	English	8 items	Dichotomous (yes/no)	
Evaluating Voice Disability–Quality of Life Questionnaire (EVD-QOL)	1996	To assess the general range of functional problems experienced with people diagnosed with voice disorders rather than to differentiate patterns of outcomes by diagnostic category	Individuals seeking evaluation for voice disorders	English	28 items in 5 subscales	5-point Likert scale (1–5)	Work Social Psychological Physical Communicative
Speech Disability Questionnaire (SDQ)	1997	To examine the effect of Botox injection on voice quality of individuals with spasmodic dysphonia and specifically on the disability arising from their voice problem	Individuals with spasmodic dysphonia	English	28 items in 5 subscales	5-point Likert scale (1–5)	Social isolation Negative communication Public avoidance Limited understanding Communication difficulty
Voice Handicap Index (VHI)	1997	To develop a psychometrically robust voice disability/handicap inventory that could be used with patients exhibiting a variety of voice disorders	Individuals with voice disorders	English	30 items in 3 subscales	5-point Likert scale (0–4)	Functional Physical Emotional
Voice Outcome Survey (VOS)	1999	To develop and validate a patient-relevant health-related quality of life instrument to evaluate vocal status of patients with uncompensated unilateral vocal fold paralysis	Individuals with uncompensated unilateral vocal fold paralysis	English	5 items	5-point Likert scale (a–e) (4 items); 3-point Likert scale (a–c) (1 item)	
Voice-Related Quality of Life (V-RQOL)	1999	To develop and validate a clinically useful instrument for measuring voice-related quality of life	Individuals with voice disorders	English	10 items in 2 subscales	5-point Likert scale (1–5)	Social-emotional physical functioning
Self-Ratings of Vocal Performance (SRVP)	1999	To assess vocal performance related to voice quality and vocal function of patients diagnosed with early glottic cancer, 6 months to 10 years after radiotherapy, compared to control speakers and to investigate consequences of voice impairment in daily life	Individuals with T1N0M0 glottic cancer	Dutch	9 items	7-point Likert scale	

(*table continues*)

Table 4. *(Continued).*

Patient-reported outcome measure	Year	Measurement aim	Target population	Language[a]	Number of items/ subscales	Response options	Subscales
Quality of Life in Recurrent Respiratory Papillomatosis (QOL-RRP)	2000	To develop a questionnaire that could be used to monitor the burden of disease on the recurrent respiratory papilloma patient in the ear, nose, and throat clinic and the speech therapy department	Individuals with recurrent respiratory papillomatosis	English	23 items	Dichotomous (yes/no)	
Voice Activity and Participation Profile (VAPP)	2001	To develop a reliable and valid tool that could be used to assess voice activity limitation and participation restriction separately	Individuals with dysphonia	Cantonese	28 items in 5 subscales	Visual analog scales (first item = *normal* to *severe;* remainder = *never* to *always*)	Self-perceived severity Job Daily communication Social communication Emotion
Voice Symptom Questionnaire (VSQ)	2003	To find out (a) how often telephone workers experience vocal symptoms, (b) how a short vocal training course affects subjective vocal symptoms, (c) relationship between change in voice symptoms and subjective effect of vocal training, and (d) how vocal training is experienced in general	Telephone customer advisors	Finnish	11 items	4-point Likert scale (1–4)	
Voice Symptom Scale (VoiSS)	2003	To devise and validate a patient-derived inventory of voice symptoms for use as a sensitive assessment tool of (a) baseline pathology and (b) response to change in adult dysphonia clinics	Individuals with dysphonia	English	43 items in 5 subscales	5-point Likert scale (1–5)	Communication problems Throat infection Psychosocial distress Voice sound and variability Phlegm
3-Item Outcome Scale (3-IOS)	2004	(a) To determine the effects of voice therapy practiced by speech therapists for patients with a chronic voice disorder and (b) to compare two self-assessment instruments in order to determine their specific utility	Individuals with chronic dysphonia	Dutch	3 items	Visual analog scales (*normal* to *extreme impairment*)	
Voice Handicap Index-10 (VHI-10)	2004	To explore and possibly develop a shortened Voice Handicap Index as a vocal function assessment tool both for initial evaluation and for longitudinal assessment of patients with voice disorders	Individuals with dysphonia	English	10 items	5-point Likert scale (0–4)	
5-Item Screening Questionnaire (5-ISQ)	2005	To assess psychometric properties of a screening questionnaire designed for detection of voice impairment in clinical practice	Individuals treated for early laryngeal cancer	Dutch	5 items	10-point Likert scale (1–10)	
Glottal Function Index (GFI)	2005	To evaluate the validity and reliability of the Glottal Function Index and to assess its utility in evaluating patients presenting with a variety of clinical entities and following treatment thereof	Individuals with vocal fold paralysis, paresis, presbylaryngis, and others	English	4 items	6-point Likert scale (0–5)	

(*table continues*)

Table 4. *(Continued).*

Patient-reported outcome measure	Year	Measurement aim	Target population	Language[a]	Number of items/ subscales	Response options	Subscales
Voice Disability Coping Questionnaire (VDCQ)	2007	To develop a disease-specific measure that would allow clinicians to focus on known symptoms and selection of items, which would contain relevant problem-focused strategies for coping with dysphonia, in a shorter form	Individuals with adductor spasmodic dysphonia and muscle tension dysphonia	English	15 items in 4 subscales	6-point Likert scale (0-5)	Social support Passive coping Avoidance Information seeking
Voice Disorder Outcome Profile (Voice-DOP)	2008	To develop a culture-specific quality of life assessment tool for individuals with voice disorders in India	Individuals with current dysphonia	Kannada	32 items in 3 subscales	Visual analog scales (*never* to *always*)	Physical Emotional Functional
Communicative Participation Item Bank (CPIB)	2009	To create an item bank to measure communicative participation across different communication disorder populations	Individuals with communication disorders	English	46 items (long); 10 items (short)	5-point Likert scale (*not at all* to *extremely*)	
Self-Evaluation of Voice as Treatment Outcome Measure (SEVTOM)	2009	To test two simple, easy-to-use questionnaires in (a) disclosing the effects of vocal loading and (b) assessing outcome of various voice hygiene interventions	Female primary school teachers	Finnish	6 items	Visual analog scale (variable)	
Scleroderma Logopedic Scale (SLS-Voice)	2010	To develop a valid and reliable tool for the assessment of oropharyngolaryngeal manifestations of scleroderma in order to obtain a quantifiable and repeatable measure	Individuals with scleroderma	Italian English	39 items in 5 subscales; 5 items in voice subscale	4-point Likert scale (1–4)	Impairment Swallow Voice Multifield Quality of life
Voice Self-Efficacy Questionnaire (VSEQ)	2011	To assess subjects' self-efficacy for voice before and after interventions	Individuals with self-declared voice problems	English	4 items	Visual analog scale (*not confident* to *extremely confident*)	
University of Rhode Island Change Assessment–Voice (URICA-Voice)	2011	To adapt the University of Rhode Island Change Assessment–Voice scale to assess the stages of readiness of patients for adherence in voice treatment	Individuals undergoing treatment for voice disorders	English Portuguese	32 items in 2 subscales	5-point Likert scale (*strongly disagree* to *strongly agree*)	Behavior Vocal use
Thyroidectomy-Related Voice Questionnaire (TVQ)	2012	To invent a simple questionnaire and evaluate its usefulness as a prethyroidectomy screening tool	Individuals scheduled to undergo thyroidectomy	English Korean	20 items	5-point Likert scale (0–4)	
Self-Efficacy in Spasmodic Dysphonia (SE-SD)	2013	To study self-efficacy in spasmodic dysphonia patients and to develop a disease-specific self-efficacy spasmodic dysphonia scale	Individuals with spasmodic dysphonia	English	11 items	4-point Likert scale (1–4)	
Screening Index for Voice Disorders (SIVD)	2013	To develop and validate a screening index for voice disorders in teachers	Female teachers	Brazilian Portuguese	12 items	4-point Likert scale (1–4)	

(*table continues*)

Table 4. *(Continued).*

Patient-reported outcome measure	Year	Measurement aim	Target population	Language[a]	Number of items/ subscales	Response options	Subscales
Work Productivity Activity Impairment Questionnaire–Specific Health Problem–Voice (WPAI-SHP)	2014	To ascertain whether existing work productivity tools are regarded by patients as adequate in assessing how the quality and quantity of a person's work is affected by spasmodic dysphonia	Individuals with spasmodic dysphonia	English	5 items	Variable	
Stanford Presenteeism Scale 6 (SPS-6)	2014	To ascertain whether work productivity tools are regarded by patients as adequate in assessing how the quality and quantity of a person's work is affected by spasmodic dysphonia	Individuals with spasmodic dysphonia	English	6 items	5-point Likert scale (1–5)	
Voice-Related Statements (VRS)	2014	To determine whether an additional set of researcher-generated voice-related statements are viewed as valuable by individuals with spasmodic dysphonia	Individuals with spasmodic dysphonia	English	14 items	5-point Likert scale (1–5)	Time Quality Quantity Personal factors
Vocal Fatigue Handicap Questionnaire (VFHQ)	2015	To construct and validate a vocal fatigue handicap questionnaire on the basis of strict convergence of distinct conceptual and psychometric criteria with the explicit goal of providing an instrument with a high degree of (a) internal consistency, (b) test–retest reliability, (c) construct and criterion validity, and (d) degree of clinical efficacy or practical relevance	Individuals with voice disorders	Italian English	30 items in 3 subscales	5-point Likert scale (0–4)	Emotional Physical Functional
Vocal Fatigue Index (VFI)	2015	To develop a psychometrically validated self-report tool (a) to generate a cohesive and consensus description of primary vocal fatigue symptoms, (b) to develop a self-report tool that can reliably and validly identify and quantify vocal fatigue symptoms, and (c) to characterize component aspects of chronic vocal fatigue	Individuals with voice complaints	English	19 items in 3 subscales	5-point Likert scale (0–4)	Tiredness/ avoidance Physical discomfort Improvement with rest

[a] Language used in initial development; does not refer to later translations.

Table 5. Patient, setting, and pathology characteristics involved in the development of voice-related patient-reported outcome measures.

Article	Patient-reported outcome measure	Study population	Setting	*N*	Distribution of pathology	Age: *M* (*SD*), range	Male	Country
Llewellyn-Thomas et al. (1984)	Linear Analog Scale Assessment of Voice Quality (LASA-VQ)	Laryngeal cancer patients who were undergoing or had undergone radiation therapy	Ontario Cancer Institute	On treatment = 30 Posttreatment = 29	Laryngeal cancer (on treatment or posttreatment) TMN: I = 21/17; II = 6/8; III = 1/3; IV = 2/1	On treatment: 60.5 (NR) Posttreatment: 60.2 (NR)	On treatment: 87.00% Posttreatment: 83.00%	Canada
Carding & Horsley (1992)	Vocal Performance Questionnaire (VPQ)	Dysphonic patients referred to the speech and language therapy clinic by staff otolaryngologists for nonorganic dysphonia	Large regional hospital	30	Nonorganic dysphonia = 30	44.3 (18.5), 18–76	23.30%	Great Britain
Mathieson (1993)	Vocal Tract Discomfort (VTD)	Patients with hyperfunctional dysphonia diagnosed at the voice clinic	Ear, nose, and throat/speech therapy voice clinic, Northwick Park Hospital	36	No mucosal changes = 12; mucosal changes = 14; other dysphonia = 10	35 (NR), 7–59	42.00%	Great Britain
Smith et al. (1996)	Evaluating Voice Disability–Quality of Life Questionnaire (EVD-QOL)	Patients from voice clinics and controls who accompanied patients seeking medical care or others who sought dental treatment or accompanied dental patients	Departments of Otolaryngology, University of Iowa and University of Utah	Cases = 174 Controls = 173	Spasmodic dysphonia = 53; neurological/ paralysis = 33; nodules = 30; laryngitis = 15; MTD = 10; bowing = 4; laryngeal trauma = 4; vocal fold scar = 3; contact ulcers = 2; miscellaneous = 20	Age / Cases, controls ≤21 / 23.6%, 8.7% 22–39 / 27.0%, 38.2% 40–65 / 27.0%, 38.2% >65 / 22.4%, 15%	Cases: 31.60% Controls: 37.20%	United States
Epstein et al. (1997)	Speech Disability Questionnaire (SDQ)	Patients with adductor spasmodic dysphonia	Middlesex Hospital outpatient department	40	Adductor spasmodic dysphonia = 40	49.6 (16.5), 20–81	42.50%	Great Britain
Jacobson et al. (1997)	Voice Handicap Index (VHI)	Patients seen in the voice clinic with a broad range of voice disorders	Voice clinic, Henry Ford Hospital	65	Mass lesion = 21; neurogenic = 17; laryngectomized = 17; MTD = 5; inflammatory = 3; atypical = 2	52.3 (16.28), NR	38.50%	United States
Gliklich et al. (1999)	Voice Outcome Survey (VOS)	Patients with unilateral uncompensated true vocal cord paralysis presenting for medialization thyroplasty and control patients presenting to emergency room with other complaints	Department of Otology and Laryngology, Harvard Medical School	Cases = 61 Controls = 48	UVFP = 61	Cases: 60.5 (18.6), 16–89 Controls: 46.0 (17.7), 21–85	Cases: 48.00% Controls: 45.80%	United States
Hogikyan & Sethuraman (1999)	Voice-Related Quality of Life (V-RQOL)	New patients presenting with a voice complaint and patients presenting for nonvoice complaints	Voice center, University of Michigan	Item refinement: 20 Validation: Cases = 109 Controls = 22	NR Inflammatory = 39; neurological = 38; mass lesions = 19; other = 13	NR Cases: 51.2 (NR), 19–85 Controls: 49.9 (NR), 19–84	NR Cases: 41.30% Controls: 40.90%	United States

(*table continues*)

Table 5. *(Continued).*

Article	Patient-reported outcome measure	Study population	Setting	*N*	Distribution of pathology	Age: *M* (*SD*), range	Male	Country
Verdonck-de Leeuw et al. (1999)	Self-Ratings of Vocal Performance (SRVP)	Patients following radiotherapy for early glottic cancer and controls with no known voice defect	Department of Otorhinolaryngology—Head and Neck Surgery, Academic Hospital Vrije Universiteit	Cases = 50 Controls = 20	T-stage: 1a = 36; 1b = 14	Age Cases, controls <65 16, 10 65–70 20, 6 70–75 8, 2 >75 6, 2	100.00%	The Netherlands
Hill et al. (2000)	Quality of Life in Recurrent Respiratory Papillomatosis (QOL-RRP)	Patients who underwent either outpatient review or surgical clearance of laryngeal papillomas	Royal National Throat, Nose, and Ear Hospital	26	RRP = 26	42.2 (NR)	53.80%	Great Britain
Ma & Yiu (2001)	Voice Activity and Participation Profile (VAPP)	Dysphonic subjects with various laryngeal pathologies	Department of Speech and Hearing Sciences, The University of Hong Kong	Item development: 45 SLP = 10 Refinement: 9 SLPs = 10 SLP students = 13 Administration: Cases = 40 Controls = 40	NR NR Nodules = 12; polyp = 3; chronic laryngitis = 9; thickened cord = 6; UVFP = 3; miscellaneous = 7	NR Cases: 41.33 (13.31), 23–58 Cases: 36.83 (10.04), 20–57 Controls: 35.65 (9.81), 20–55	NR Cases: 11.00% Cases: 20.00% Controls: 20.00%	Hongkong
Lehto et al. (2003)	Voice Symptom Questionnaire (VSQ)	Customer advisors who mainly use the telephone during their work hours at a call center	Telecommunications operator Sonera	48	Edema = 3; erythema = 7; incomplete closure = 1; normal = 37	Female: 29 (NR), 21–40 Male: 26 (NR), 21–38	20.80%	Finland
Scott et al. (1997)	Voice Symptom Scale (VoiSS), Phase 1	Patients referred complaining of hoarseness	Phoniatric clinic, Glasgow Royal Infirmary	133	NR	54 (NR), 18–80	32.30%	Great Britain
Deary et al. (2003)	Voice Symptom Scale (VoiSS), Phases 2 and 3	Typical patients with dysphonia presenting to the ear, nose, and throat department	Voice center, University of Newcastle	Pilot = 168 Refinement = 180	NR Functional = 51; vocal cord palsy = 25; laryngitis = 21; "acid" laryngitis = 9; Reinke's edema = 9; asthma = 6; malignancy = 6; RRP = 6; globus/phlegm = 5; nodules = 5; leukoplakia = 4; polyp = 3; granuloma = 3; exophytic lesion = 2; cricoarytenoid joint arthritis = 2; puberphonia = 2; miscellaneous = 6; resolving/normal = 15	Male: 49.8 (16.0) Female: 48.4 (13.9) Male: 55.4 (14.0) Female: 53.4 (16.0)	25.60% 35.00%	Great Britain

(*table continues*)

Table 5. *(Continued).*

Article	Patient-reported outcome measure	Study population	Setting	*N*	Distribution of pathology	Age: *M* (*SD*), range	Male	Country

Speyer et al. (2004)	3-Item Outcome Scale (3-IOS)	Patients with chronic dysphonia diagnosed by phoniatrician	Phoniatric department, University Hospital Utrecht	77	MTD = 12; submucosal swelling = 7;
					nodules = 10;
					polyps = 6;
					UVFP = 7;
					slight vocal fold abnormalities = 24;
					severe vocal fold abnormalities = 11
						Male: 47 (NR)
						Female: 40 (NR)
						Range: 18–76
							44.20%	The Netherlands

Rosen et al. (2004)	Voice Handicap Index-10 (VHI-10)	Patients with dysphonia presenting to the voice clinic and volunteers consisting of family members of patients visiting the department	Laryngology clinic, University of Pittsburgh	Item analysis:
				Cases = 100
				Controls = 159
				Longitudinal = 59
				Comparative:
				Cases = 819
				Controls = 173
					Functional = 5;
					allergic laryngitis = 2;
					LPR = 13; MTD = 11;
					neurologic, other = 2;
					Parkinson's = 2;
					paradoxical vocal fold motion = 1;
					psychologic disease = 1;
					puberphonia = 1;
					Reinke's edema = 7;
					RRP = 3;
					abductor spasmodic dysphonia = 1;
					adductor spasmodic dysphonia = 2;
					subglottic stenosis = 1;
					atrophy = 3; cancer = 1;
					cyst = 10; vocal process granuloma = 1;
					nodules = 7; UVFP = 10;
					paresis = 1; polyp = 9;
					scar = 5
					Functional = 1;
					MTD = 13; paradoxical vocal fold motion = 3;
					Reinke's edema = 3;
					cancer = 1; cyst = 4;
					granuloma = 1;
					nodules = 2; UVFP = 15;
					paresis = 2; polyp = 10;
					scar = 4
					UVFP = 104;
					paresis/atrophy = 90;
					MTD = 147;
					polyp/cyst/nodule = 166;
					paradoxical vocal fold motion = 87;
					functional = 32;
					reflux = 37; scar = 54;
					neurologic = 52;
					Reinke's edema = 27;
					adductor spasmodic dysphonia = 23
						NR
						NR
						NR
							NR
							NR
							NR
								United States

(*table continues*)

Table 5. *(Continued).*

Article	Patient-reported outcome measure	Study population	Setting	*N*	Distribution of pathology	Age: *M* (*SD*), range	Male	Country
Arffa et al. (2012)	Voice Handicap Index-10 (VHI-10), normative	Family members of otolaryngology patients without voice complaints	Laryngology clinic, University of Pittsburgh	156	Healthy controls = 156	NR	32.00%	United States
van Gogh et al. (2005)	5-Item Screening Questionnaire (5-ISQ)	Patients visiting the outpatient clinic for follow-up visit after initial radiation or endoscopic surgery for early glottic cancer and controls without voice complaints	Department of Otorhinolaryngology—Head and Neck Surgery, Vrije University Medical Center	Cases = 177 (radiation = 126; surgery = 51) Controls = 110	Laryngeal cancer (dysplasia to T2)	Cases: Radiation = 66 (NR), 39–80 Surgery = 66 (NR), 40–81 Controls: 61 (NR), 40–80	Cases: Radiation = 92.90% Surgery = 88.20% Controls = 50.00%	The Netherlands
Bach et al. (2005)	Glottal Function Index (GFI)	Patients presenting with dysphonia resulting from a variety of clinical entities and following treatment thereof	Center for Voice Disorders, Wake Forest University	Item development = NR Responsiveness = 40 Specificity: Cases = 120 Controls = 40	NA Glottic insufficiency = 40 Nodules = 40; adductor spasmodic dysphonia = 40; granuloma = 40	NA Median: 49 Cases: NR Controls: median = 39	NA 37.00% Cases: NR Controls: 50.00%	United States
Laukkanen et al. (2009)	Self-Evaluation of Voice as Treatment Outcome Measure (SEVTOM)	Female primary school teachers with functionally healthy voices	Recruited via Internet questionnaire	90	Not evaluated/ functionally normal = 90	41.1 (8.5)	0.00%	Finland
Baylor et al. (2009)	Communicative Participation Item Bank (CPIB)	Adults with spasmodic dysphonia	Recruited through multiple sources: NSDA, ASHA SIG 3, and local voice clinics	208	Spasmodic dysphonia = 208	55.4 (11.0), 27–83	22.10%	United States
Baylor et al. (2013)	Communicative Participation Item Bank (CPIB)	Adults with multiple sclerosis, Parkinson's disease, amyotrophic lateral sclerosis, and head and neck cancer	Recruited through multiple sources: Internet listserv, support groups, and local disease registries	701	Multiple sclerosis = 216; Parkinson's disease = 218; amyotrophic lateral sclerosis = 70; head and neck cancer = 197	58.8 (12.4), 24–99	45.70%	United States
Epstein et al. (2009)	Voice Disability Coping Questionnaire (VDCQ)	Voice clinic referrals at one hospital	Voice clinic, Royal National Throat, Nose, and Ear Hospital and Ear Institute	80	Adductor spasmodic dysphonia = 40; MTD = 40	Overall: 45.4 (NR) Adductor spasmodic dysphonia: 49.70 (16.281) MTD: 41.31 (19.569)	35.00%	Great Britain
Konnai et al. (2010)	Voice Disorder Outcome Profile (Voice-DOP)	Individuals with current dysphonia and age-matched controls	All India Institute of Speech and Hearing, Mysore; St. Johns Medical College and Hospital, Bangalore; Government ENT Hospital, Hyderabad	Development: SLPs = 10 SLP students = 10 People with dysphonia = 5 Refinement: SLPs = 10 SLP students = 5 Administration: Cases = 42 Controls = 30	Different vocal pathologies = 5 NA Nodules = 6; glottic chink = 6; carcinoma = 6; gastroesophageal reflux disease = 6; puberphonia = 7; UVFP = 4; laryngitis = 2; atypical = 5	NR NA 3. 34 (NR), 18–60	NR NA 3.83%	India

(*table continues*)

Table 5. *(Continued).*

Article	Patient-reported outcome measure	Study population	Setting	*N*	Distribution of pathology	Age: *M* (*SD*), range	Male	Country
Vitali et al. (2010)	Scleroderma Logopedic Scale (SLS-Voice)	Patients with laboratory-defined systemic sclerosis disease	Ospedale Maggiore, Immunology Clinical Unit	Focus group = NR Pilot = 28 Phase 2 = 16 Phase 3 = 15 Administration: Cases = 86 Controls = 40	Systemic sclerosis *N* = 28 *N* = 16 *N* = 15 *N* = 86	NR NR NR Cases: 57.0 (12.7) Controls: 56.1 (7.5)	NR NR NR Cases: 19.00% Controls: 20.00%	Italy
Gillespie & Abbott (2011)	Voice Self-Efficacy Questionnaire (VSEQ)	Teachers from the Pittsburgh Public School District who reported current or past self-identified voice problem	University of Pittsburgh Voice Center	Item development: SLPs = 2 Administration = 14	NR	NR	14.00%	United States
Teixeira et al. (2013)	University of Rhode Island Change Assessment–Voice (URICA-Voice)	Patients receiving voice therapy at the university hospital at two institutions	Department of Speech-Language Pathology and Audiology, Hospital of the Universidade Federal de Minas Gerais and Hospital San Paulo, Universidade Federal de Sao Paulo	66	Behavioral dysphonia = 60 Nonbehavioral dysphonia = 6	42.27 (NR), 18–68	12.10%	Brazil

Nam et al. (2012)	Thyroidectomy-Related Voice Questionnaire (TVQ)	Patients scheduled to undergo thyroidectomy	Department of Otolaryngology and Surgery, The Catholic University of Korea	500	LPR = 136;
					nodules = 24;
					polyp = 9;
					vocal fold palsy = 6;
					Reinke's edema = 2;
					cyst = 1;
					vocal sulcus = 1;
					normal = 321
						45.5 (11.97), 16–76	17.60%	Korea

Hu et al. (2013)	Self Efficacy in Spasmodic Dysphonia (SE-SD)	Spasmodic dysphonia patients	Department of Otolaryngology, University of Washington	Item development: Laryngologists = 3 Fellow = 1 SLP = 3 Patients = 2 Administration = 145	NR Adductor spasmodic dysphonia = 139; abductor spasmodic dysphonia = 6	NR 59.5 (13.6)	NR 24.80%	United States
Ghirardi et al. (2013)	Screening Index for Voice Disorders (SIVD)	Current female teachers from the public school system in San Paulo, Brazil	Pontifical Catholic University of San Paulo and School of Public Health at University of San Paulo	Item development = 252 Internal validation = 130 External validation = 122	NR NR Voice disorder = 73; no voice disorder = 49	NR 40.6 (NR) 39.4 (NR)	0.00%	Brazil
Isetti & Meyer (2014)	Work Productivity Activity Impairment Questionnaire–Specific Health Problem–Voice (WPAI-SHP)	Patients diagnosed with spasmodic dysphonia getting treatment at the academic voice center	Department of Otolaryngology, University of Washington	9	Spasmodic dysphonia = 9	51 (13.2), 28–71	33.00%	United States
Isetti & Meyer (2014)	Stanford Presenteeism Scale 6 (SPS-6)	Patients diagnosed with spasmodic dysphonia getting treatment at the academic voice center	Department of Otolaryngology, University of Washington	9	Spasmodic dysphonia = 9	51 (13.2), 28–71	33.00%	United States

(*table continues*)

Table 5. *(Continued).*

Article	Patient-reported outcome measure	Study population	Setting	*N*	Distribution of pathology	Age: *M* (*SD*), range	Male	Country
Isetti & Meyer (2014)	Voice-Related Statements (VRS)	Patients diagnosed with spasmodic dysphonia getting treatment at the academic voice center	Department of Otolaryngology, University of Washington	9	Spasmodic dysphonia = 9	51 (13.2), 28–71	33.00%	United States
Paolillo & Pantaleo, (2015)	Vocal Fatigue Handicap Questionnaire (VFHQ)	Patients with voice disorders at an academic medical center	San Leopoldo Mandic Hospital	Item development: Laryngologists = 3 SLP = 1 Patients = 20 Item reduction = 30 Validation = 87	Neurogenic = 2; structural = 9; inflammatory = 4; functional = 5 Neurogenic = 2; structural = 15; inflammatory = 7; functional = 6 Neurogenic = 6; structural = 44; inflammatory = 19; functional = 18	44.28 (15.04), NR 43.82 (12.4), NR 43.33 (16.83), NR	40.00% 30.00% 28.00%	Italy
Nanjundeswaran et al. (2015)	Vocal Fatigue Index (VFI)	Voice clinic referrals at two academic voice centers	University of Pittsburgh Voice Clinic, Vanderbilt Voice Center	Item development: Laryngologists = 4 SLPs = 6 Initial testing = 197 Validation: Cases = 98 Controls = 70	NA Atrophy = 29; membranous lesion = 46; MTD/dysphonia = 48; granuloma = 3; paralysis = 18; scar = 6; ADSD = 16; laryngitis = 4; cancer = 4; tremor = 2; RRP = 3; l eukoplakia = 9; paresis = 1; LPR = 1; hemorrhage = 1; chondroma = 1; NOS = 3 Scar = 6; paralysis = 14; atrophy = 15; paresis = 7; membranous lesion = 24; granuloma = 1; MTD/dysphonia = 17; RRP = 2; LPR = 2; edema = 2; hemorrhage = 2; subepithelial mass = 1; leukoplakia = 3; laryngitis = 1; NOS = 8	NA UPVC: 51.76 (19.54) VVC: 50.94 (16.01) Controls: 39 (15) Cases: NR	NA 37.00% Controls: 30.00% Cases: NR	United States

Note. TNM =cancer stage (T = tumor size, N = nodal status, M = metastases); NR = not recorded; MTD = muscle tension dysphonia; UVFP = unilateral vocal fold paralysis; SLP = speech-language pathologist; LPR = laryngopharyngeal reflux; RRP = recurrent respiratory papillomatosis; NSDA = National Spasmodic Dysphonia Association; ASHA = American Speech-Language-Hearing Association; SIG 3 = American Speech-Language-Hearing Association Special Interest Group 3; ADSD = adductor spasmodic dysphonia; NOS = not otherwise specified; UPVC = University of Pittsburgh Voice Center; VVC = Vanderbilt Voice Center.

Voice-Related Statements (VRS; Isetti & Meyer, 2014)

- Activity limitation: VAPP
- Disability: Speech Disability Questionnaire (SDQ; Epstein, Stygall, & Newman, 1997)
- Burden of disease: Quality of Life in Recurrent Respiratory Papillomatosis (QOL-RRP; Hill, Akhtar, Corroll, & Croft, 2000)
- Voice symptoms: GFI, Voice Symptom Scale (VoiSS; Deary, Wilson, Carding, & MacKenzie, 2003; Scott, Robinson, Wilson, & Mackenzie, 1997), Voice Symptom Questionnaire (VSQ; Lehto, Rantala, Vilkman, Alku, & Backstrom, 2003), SLS-voice
- Vocal tract discomfort: Vocal Tract Discomfort (VTD; Mathieson, 1993)
- Adherence to voice therapy: University of Rhode Island Change Assessment–Voice (URICA-Voice; Teixeira et al., 2013)
- Screening for voice disorders: SIVD, 5-Item Screening Questionnaire (5-ISQ; van Gogh et al., 2005)

Disease-Specific and General Voice-Related PRO Measures

Instruments were divided into condition-specific and general voice PRO measures (see Table 4). Diseases or conditions specifically targeted included laryngeal cancer (LASA-VQ, 5-ISQ, SRVP), nonorganic or hyperfunctional dysphonia (VPQ), spasmodic dysphonia (VDCQ, SDQ, SE-SD, WPAI-SHP/SPS-6/VRS), unilateral vocal fold paralysis (VOS), recurrent respiratory papillomatosis (QOL-RRP), scleroderma (SLS-Voice), and patients undergoing thyroidectomy (TVQ). Three focused on occupational voice use, including two specific to teachers (SEVTOM, SIVD) and one for telephone customer advisors (VSQ). The remaining PRO measures were aimed at a more general respondent population that may either be at risk for or currently have various voice-related or communication-impairing conditions.

Assessment of Measurement Characteristics

Figure 4 provides an itemized, schematic overview of the measurement characteristics and utility for the 32 identified voice-related PRO measures. Predominant patterns among these instruments with domain-specific examples are described below. The VHI and VHI-10 are the most commonly used voice-related PRO measures and the most familiar to practitioners and clinical researchers. Because of their familiarity and common use, we chose to use these instruments as exemplars when possible.

Conceptual Model

Development of the PRO measures varied. All included some description of the respective conceptual model. Each defined the PRO construct to be measured and identified the intended respondent population. Deficiencies related to failure to address whether the tool was expected to have a single scale or multiple subscales were noted in four of 32 PRO measures: VPQ, VTD, VSQ, and VSEQ (see Figure 4).

Content Validity

All measures demonstrated some degree of content validity, albeit with considerable inconsistency and variable methodological rigor. Most (22/32) provided some description of the process used to derive included items in their respective PRO measure (see Figure 4). In contrast, few PRO measures included subjects in the item development process. Ten (41%) sought direct patient input (e.g., focus groups, interviews) to inform the content of items included in the respective measures: LASA-VQ, VOS, V-RQOL, VAPP, VoiSS, CPIB, Voice-DOP, SLS-Voice, WPAI-SHP/SPS-6/VRS, and VFHQ. In contrast, all 32 included content experts (e.g., voice practitioners, speech-language pathologists, laryngologists, otolaryngologists) in the instrument construction process.

Reliability

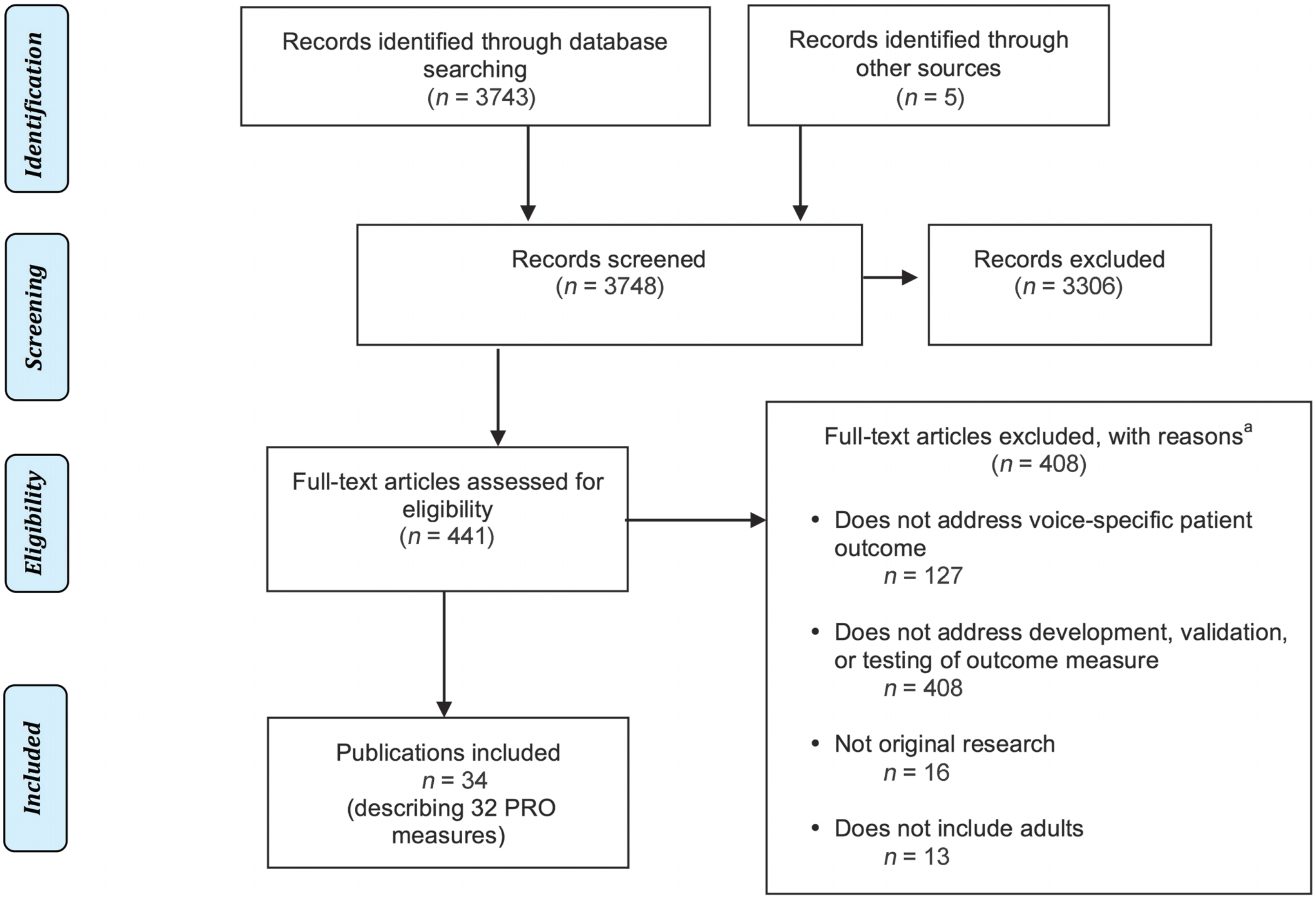
Reliability (e.g., test–retest, internal consistency) of the final proposed PRO measure was not tested in nine of 32 identified: VTD, EVD-QOL, SRVP, VSQ, VHI-10, GFI, URICA-Voice, TVQ, and WPAI-SHP/SPS-6/VRS (see Figure 4). Some studies performed reliability testing, but not on the final PRO measure items. For instance, in developing the VHI-10, initial item-analysis reliability testing identified nine items (VHI-9) that had combined coefficient alpha of 0.90. After reliability testing, a clinical consensus conference of content experts convened to determine which of the original 30 VHI items were most clinically relevant. Only five items identified by the experts overlapped with the reliability-tested VHI-9. Thus, five reliability-tested items were combined with five items deemed most clinically relevant. The final combined 10 items in the VHI-10 were not retested for reliability. Among the PRO measures that appropriately tested reliability, 20 of 21 reported adequate reliability indices ($r \geq .70$) or justified lower values (Aaronson et al., 2002; Reeve et al., 2013).

Construct Validity

The 32 studies included were less consistent in demonstrating various forms of construct validity, one of which evaluated whether a measure's factor or subscale structure was empirically justified. Overall, 12 of 32 voice-related PRO measures statistically verified the existence of either a single scale (e.g., common factor) or discrete subscales: SDQ, VOS, QOL-RRP, VoiSS, VHI-10, 5-ISQ, CPIB, VDCQ, SLS-Voice, SE-SD, SIVD, and VFI (see Figure 4). Others, including the VHI, did not provide a statistical basis for their subscale structure (i.e., emotional, functional, and physical).

A second characteristic considered in this category was longitudinal validity, defined as both demonstrated test–retest reliability (e.g., stability in the absence of any known change) and responsiveness to change. This criterion

Figure 3. Disposition of studies identified for this review. PRO = patient-reported outcome.


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 3743)
Records identified through other sources (n = 5)
Records screened (n = 3748)
Records excluded (n = 3306)
Full-text articles assessed for eligibility (n = 441)
Full-text articles excluded, with reasons[a] (n = 408)
• Does not address voice-specific patient outcome n = 127
• Does not address development, validation, or testing of outcome measure n = 408
• Not original research n = 16
• Does not include adults n = 13
Publications included n = 34 (describing 32 PRO measures)

[a] Numbers do not tally as studies could be excluded for multiple reasons.

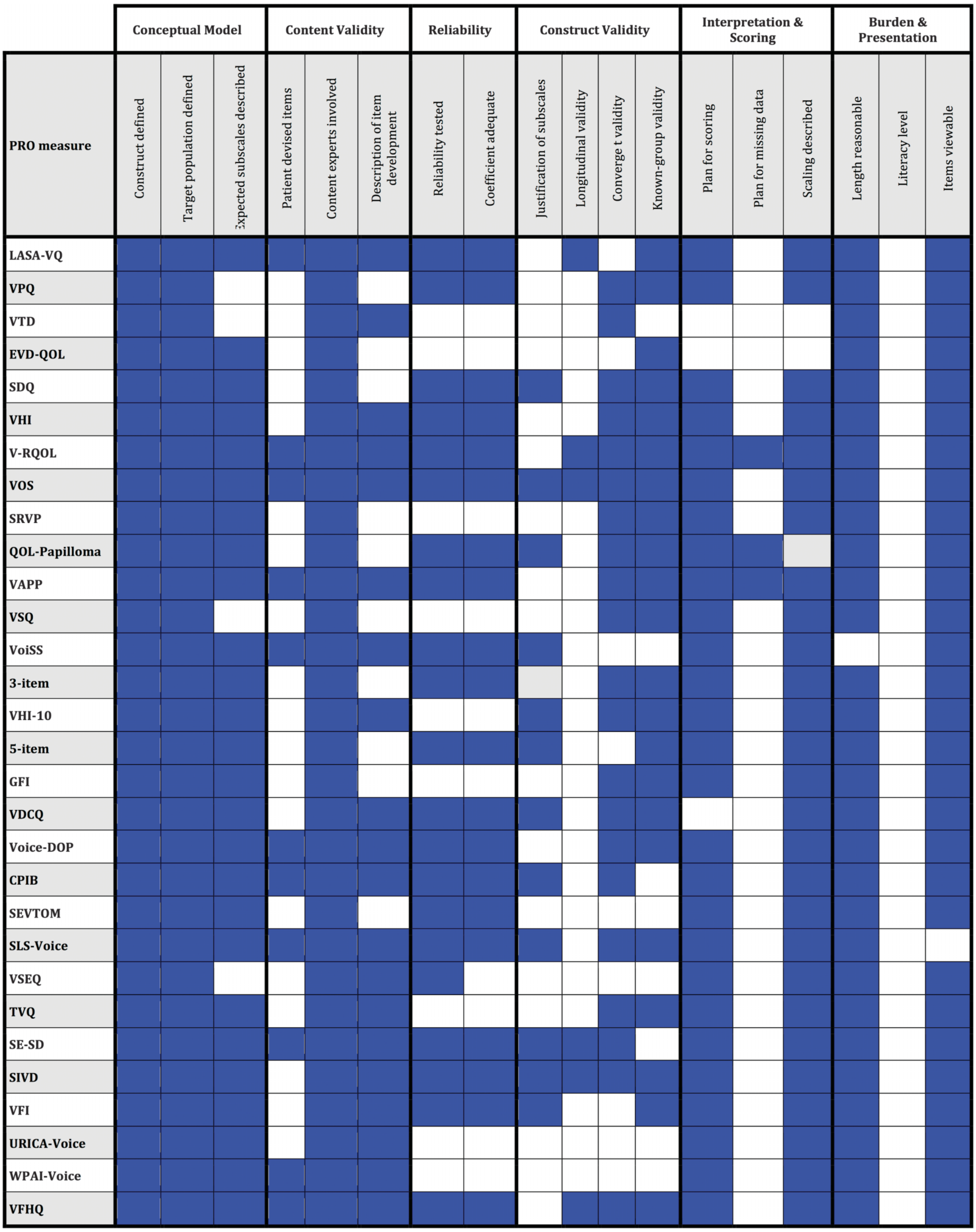
was alternatively met if the PRO measure was designed for screening purposes and this was explicitly stated and/or the measure was not intended to track change in its construct over time. Only three studies adequately demonstrated evidence of longitudinal validity (VOS, V-RQOL, LASA-VQ), and an additional three fulfilled the criterion as they were designed for screening purposes and/or did not intend at the time of publication to track their construct's change over time (SE-SD, SIVD, VFHQ).

Convergent validity is a third form of construct validity evaluated. It exists when the candidate PRO measure shows an expected association with other existing PRO measures that focus on similar construct or relevant clinical correlates (e.g., objective, physiologic data). Overall, 21 of 32 studies showed evidence of this form of validity: VPQ, VTD, SDQ, VHI, VOS, V-RQOL, SRVP, QOL-RRP, VAPP, VSQ, 3-IOS, VHI-10, GFI, CPIB, VDCQ, Voice-DOP, SLS-Voice, TVQ, SE-SD, SIVD, and VFHQ (see Figure 4). A fourth form of construct validity, known-group validity, refers to expected differences between groups known to differ in the degree of the construct (e.g., cases and controls). In all, 22 of 32 PRO measures met this criterion: LASA-VQ, VPQ, EVD-QOL, SDQ, VHI, VOS, V-RQOL, SRVP, QOL-RRP, VAPP, VSQ, 3-IOS, VHI-10, 5-ISQ, GFI, VDCQ, Voice-DOP, SLS-Voice, TVQ, SIVD, VFHQ, and VFI (see Figure 4).

Scoring and Interpretation

Each study provided some description of either how to score the measure or a means to interpret the scores. Although the majority provided both, three PRO measures did not describe the proposed scoring approach or algorithm: VDCQ, VTD, and EVD-QOL. Among those that did, the most common scoring method was simple summation. No PRO measure addressed missing data. In other words, none offered a plan for managing or interpreting PRO measures that had missing responses. A third parameter assessed whether an explanation of how to interpret the PRO

Figure 4. Summary comparison of measurement properties among identified patient-reported outcome (PRO) measures. LASA-VQ = Linear Analog Scale Assessment of Voice Quality; VPQ = Vocal Performance Questionnaire; VTD = Vocal Tract Discomfort; EVD-QOL = Evaluating Voice Disability–Quality of Life Questionnaire; SDQ = Speech Disability Questionnaire; VHI = Voice Handicap Index; V-RQOL = Voice-Related Quality of Life; VOS = Voice Outcome Survey; SRVP = Self-Ratings of Vocal Performance; QOL-RRP = Quality of Life in Recurrent Respiratory Papillomatosis; VAPP = Voice Activity and Participation Profile; VSQ = Voice Symptom Questionnaire; VoiSS = Voice Symptom Scale; 3-IOS = 3-Item Outcome Scale; VHI-10 = Voice Handicap Index-10; 5-ISQ = 5-Item Screening Questionnaire; GFI = Glottal Function Index; VDCQ = Voice Disability Coping Questionnaire; Voice-DOP = Voice Disorder Outcome Profile; CPIB = Communicative Participation Item Bank; SEVTOM = Self-Evaluation of Voice as Treatment Outcome Measure; SLS-Voice = Scleroderma Logopedic Scale; VSEQ = Voice Self-Efficacy Questionnaire; TVQ = Thyroidectomy-Related Voice Questionnaire; SE-SD = Self-Efficacy in Spasmodic Dysphonia; SIVD = Screening Index for Voice Disorders; VFI = Vocal Fatigue Index; URICA-Voice = University of Rhode Island Change Assessment–Voice; WPAI-SHP = Work Productivity Activity Impairment Questionnaire–Specific Health Problem–Voice; VFHQ = Vocal Fatigue Handicap Questionnaire.



	Conceptual Model			Content Validity			Reliability		Construct Validity				Interpretation & Scoring			Burden & Presentation		
PRO measure	Construct defined	Target population defined	Expected subscales described	Patient devised items	Content experts involved	Description of item development	Reliability tested	Coefficient adequate	Justification of subscales	Longitudinal validity	Converge t validity	Known-group validity	Plan for scoring	Plan for missing data	Scaling described	Length reasonable	Literacy level	Items viewable
LASA-VQ	■	■	■	■	■	■	■	■	□	■	□	■	■	□	■	■	□	■
VPQ	■	■	□	□	■	□	■	■	□	□	■	■	■	□	■	■	□	■
VTD	■	■	□	□	■	■	□	□	□	□	■	□	□	□	□	■	□	■
EVD-QOL	■	■	■	□	■	□	□	□	□	□	□	■	□	□	□	■	□	■
SDQ	■	■	■	□	■	□	■	■	■	□	■	■	■	□	■	■	□	■
VHI	■	■	■	□	■	■	■	■	□	□	■	■	■	□	■	■	□	■
V-RQOL	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	□	■
VOS	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	□	■
SRVP	■	■	■	□	■	□	□	□	□	□	■	■	■	□	■	■	□	■
QOL-Papilloma	■	■	■	□	■	□	■	■	■	□	■	■	■	■	▨	■	□	■
VAPP	■	■	■	■	■	■	■	■	□	□	■	■	■	■	■	■	□	■
VSQ	■	■	□	□	■	□	□	□	□	□	■	■	■	□	■	■	□	■
VoiSS	■	■	■	■	■	■	■	■	■	□	□	□	■	□	■	□	□	■
3-item	■	■	■	□	■	□	■	■	▨	□	■	■	■	□	■	■	□	■
VHI-10	■	■	■	□	■	■	□	□	■	□	■	■	■	□	■	■	□	■
5-item	■	■	■	□	■	□	■	■	■	□	□	■	■	□	■	■	□	■
GFI	■	■	■	□	■	□	□	□	□	□	■	■	■	□	■	■	□	■
VDCQ	■	■	■	□	■	■	■	■	■	□	■	■	□	□	■	■	□	■
Voice-DOP	■	■	■	■	■	■	■	■	□	□	■	■	■	□	■	■	□	■
CPIB	■	■	■	■	■	■	■	■	■	□	■	□	■	□	■	■	□	■
SEVTOM	■	■	■	□	■	□	■	■	□	□	□	□	■	□	■	■	□	■
SLS-Voice	■	■	■	■	■	■	■	■	■	□	■	■	■	□	■	■	□	□
VSEQ	■	■	□	□	■	■	■	□	□	□	□	□	■	□	■	■	□	■
TVQ	■	■	■	□	■	■	□	□	□	□	■	■	■	□	■	■	□	■
SE-SD	■	■	■	■	■	■	■	■	■	■	■	□	■	□	■	■	□	■
SIVD	■	■	■	□	■	■	■	■	■	■	■	■	■	□	■	■	□	■
VFI	■	■	■	□	■	■	■	■	■	□	□	■	■	□	■	■	□	■
URICA-Voice	■	■	■	□	■	■	□	□	□	□	□	□	■	□	■	■	□	■
WPAI-Voice	■	■	■	■	■	■	□	□	□	□	□	□	■	□	■	■	□	■
VFHQ	■	■	■	■	■	■	■	■	□	■	■	■	■	□	■	■	□	■

measure score was provided. Most (29/32) stated that higher scores were worse without offering mathematically justified scaling (see Figure 4). In fact, no PRO measures offered statistically justified anchors to help interpret what a particular severity score indicated.

Respondent Burden and Presentation

Current reviewers felt the time needed to complete the items was reasonable in 31 of 32 voice-related PRO measures (see Figure 4). All but one instrument presented their full set of questions in a published article or referenced an accessible source (Vitali et al., 2010). No study described the literacy level for their respective PRO measure.

Discussion

Growing emphasis on patient-centered outcomes research and comparative effectiveness research has increased the need to capture accurate, patient-centered data to evaluate the effectiveness of treatment, management, and direct decision making. PRO measures are the predominant method used to systematically collect patient perspective and experience. They can be designed to quantify relatively qualitative phenomena for which objective measures are lacking or inadequate. This task is particularly applicable and relevant to voice disorders wherein the patient's perspective contextualizes its severity on disability or quality of life.

Recognition of the importance of PRO measures in voice disorders has sparked rapid growth in their number and construct diversity (e.g., quality of life, coping) since the first PRO measure was introduced into this field in 1984 (LASA-VQ). Questionnaires also differ in applicability. Some voice-related PRO measures are disease specific (e.g., spasmodic dysphonia, unilateral vocal fold paralysis), whereas others are designed to assess a broader, diverse population. This expansion unfortunately has also led to wide variability in the methodological rigor of their development. Identifying the appropriate PRO measure for a given purpose requires nuanced understanding of a particular measure's underlying conceptual model and methodological properties (e.g., clinimetrics, psychometrics).

This systematic review extracted existing voice-related PRO measures and reviewed their content, measurement characteristics, and applicability. A total of 32 voice-related instruments were identified and analyzed. Catchment was higher than prior systematic reviews (Biddle et al., 2002; Branski et al., 2010; Franic et al., 2005) on the basis of the current review's contemporariness, broad inclusion criteria, and exhaustive systematic literature search.

As expected, psychometric rigor was quite disparate among identified PRO measures. The range of target individuals involved in the development and/or validation of these measures varied significantly from nine to more than 1,300 (Isetti & Meyer, 2014; Rosen et al., 2004). It is generally recommended that variable and subject sampling are optimized for factor/principal components analysis–based methods and/or that there be more than 100 subjects involved in validation (Terwee et al., 2007; Velicer & Fava, 1998). Few studies achieved this standard. Adequacy and applicability of measures that include few individuals from the target population in development should be questioned. Three PRO measures identified were developed for use in another population and were being adapted and tested post hoc for application in individuals with voice problems (Isetti & Meyer, 2014; Teixeira et al., 2013). Revalidating a measure before applying it to a new target population is an often overlooked and important step. Distorted results are a risk when end users implement implement PRO measures designed for a specific target population into a new application or population.. Outcomes that are based on responses to a measure that is not designed for a particular population may not be valid. Thus, it is important for PRO measure developers to fully describe the population used in their creation (e.g., age, gender, race, disorder). Many voice-related tools did not provide this information.

Lack of Patient Centeredness

Only 41% of PRO measures directly engaged patients in the item development stage (i.e., devising items) despite claiming to be patient centric. Lack of empiric content data from patients significantly limits content validity. The foundation for PRO measures is the target population perspective and experience. Thus, omitting patients at this stage compromises the validity of scores and creates a condition in which patients answer questions that are designed by and based on the experience and opinions of content experts who do not live with their particular condition. This concern was highlighted in a prior systematic review regarding content validity of voice-related PRO measures (Branski et al., 2010). The report found that five of nine PRO measures did not include patients in item development. It is interesting to note that the VHI was one measure found to have included patients in this process. Careful re-evaluation of the original VHI article and correspondence with authors reveals that items were derived from case history in chart reviews, not from empiric patient interviews or focus groups. Case histories rely on provider recollection and therefore are a provider's interpretation of the patient's perspective. This deficiency was not recognized by prior systematic reviews (Biddle et al., 2002; Branski et al., 2010; Franic et al., 2005). As a consequence, the content validity of the VHI and, transitively, the VHI-10 may be overestimated.

Development Characteristics

This analysis was not designed to evaluate every study that used or translated voice-related PRO measures. Rather, it focused on identifying and evaluating the strengths and weaknesses of measurement properties. Most published measures purported reliability and validity. This simple statement is often considered sufficient legitimization of a PRO measure's quality by end users. It is important to recognize that reliability and validity are not discrete concepts and exist on a spectrum (Newton & Shaw, 2013).

Half of voice PRO measures (50%) met at least one criteria in each measurement domain assessed, and none met all criteria. Reasons for deficiencies are multifactorial. Some may be due to lack of understanding of the complex underlying psychometric methodology and the time intensity necessary to create a high-quality PRO measure. As an alternative, other instruments still may be under development. Often, PRO measures are devised and published in stages, as exemplified by the VoiSS, VHI-10, and CPIB.

Construct Validity

PRO measure development was scrutinized for quantitative justification of proposed subscales such as factor analysis or item response theory techniques (Hambleton & Swaminathan, 1985). The VHI and 13 other instruments cited subscales (e.g., emotional, physical, and functional) without evaluating their empirical basis using item-level analysis (see Figure 4). Without proven, discrete subscales, all items may measure the same overall construct. For instance, an analysis of VHI items during the development of the VHI-10 found a common factor: All VHI questions, despite assignment to the social, emotional, or functional subscales, measure aspects of the same construct. However, a separate factor analysis performed as part of cross-cultural adaption of the VHI and VHI-10 was able to demonstrate the VHI's three-factor solution (Lam et al., 2006). In contrast, two studies evaluating the factor structure (one using Rasch analysis) found only two unidimensional constructs or factors (Bogaardt, Hakkesteegt, Grolman, & Lindeboom, 2007; Wilson et al., 2004). Thus, controversy remains over what discrete construct(s) are empirically measured by the VHI. This issue may also exist for the V-RQOL and other measures that use distinct subscales without statistical justification.

We also evaluated responsiveness to change (longitudinal validity). Instruments that aim to measure change over time should demonstrate stability in score when no change is expected. Otherwise, distinguishing "real" from random or "chance" differences becomes difficult. Test–retest reliability is used to evaluate score stability. Scores should also show meaningful change in an expected direction after an intervention (responsiveness to change). This property is a claimed attribute of most voice-related PRO measures; that is, they should be able to track change in the construct (e.g., dysphonia severity) over time. However, in this analysis, only three of 32 instruments met the criterion: VOS, V-RQOL, and LASA-VQ. Some measures have demonstrated responsiveness to change in subsequent studies since their initial development (e.g., VHI, VPQ). Nonetheless, on the basis of their initial development, many voice-related PRO measures may not be appropriate for and give spurious results in clinical trials and other comparative effectiveness studies.

Two other forms of construct validity evaluated in this review were convergent validity and known-group validity. The former relates to whether the proposed PRO measure correlates in an expected way with either an existing PRO measure(s) or clinical data that quantify the same concept. The latter is the ability of the instrument to distinguish among those respondents who are expected to differ. Most voice-related PRO measures (25/32) met at least one of these assessed variants of construct validity.

Scoring and Interpretation

Most voice-related PRO measures (26/32) provided some degree of scoring instructions. Most also provided some information on score interpretation, typically indicating that higher scores correlate with greater disease burden. Interpreting scores derived from summation is a common problem faced by end users. It is not always clear what a score means or when a minimally important clinical difference exists (Guyatt et al., 2002). A clinically important change should be a change in score that patients consider to be important. Many strategies have been proposed and depend on whether the intended use is within-person or population-based analysis (Guyatt et al., 2002). Estimation of this value for a particular measure is important to its interpretability. Its omission in the validation process represents a critical weakness in currently available PRO measures and limits their usefulness in decision making and implementation in clinical trials.

Incomplete questionnaires are also common occurrences in both clinical practice and research applications. Implications of missing PRO measure data can be significant, particularly if answers are missing systematically, thus introducing bias. For example, incomplete data could result from the reading level of the instrument or to questions that may not pertain to all respondents (e.g., occupational questions in retired individuals). A strategy for accommodating missing data is important to consider when developing a PRO measure. Many techniques for dealing with incomplete data exist. However, in contrast to scoring information, only three of the 32 identified PRO measures mentioned management of incompletely scored questionnaires. In fact, the only strategy cited was to omit incomplete questionnaires (Hill et al., 2000; Hogikyan & Sethuraman, 1999; Ma & Yiu, 2001).

Respondent Burden and Presentation

PRO measures inherently place burden on respondents in terms of length and complexity. Reasonable length depends on the intended respondent population and the setting in which it is being administered (e.g., clinic vs. research laboratory). With few exceptions, identified PRO measures were considered to have reasonable time to complete, ranging in length from three to 32 items.

No instrument described its literacy level. Some instruments identified in the current study overlapped with those identified in a prior review that focused on readability among voice-related PRO measures (Zraick & Atcherson, 2012). In all, the seven overlapping PRO measures had Flesch-Kincaid Reading Ease scores (Kincaid, Fishburne, Rogers, & Chissom, 1975) ranging from 89 (fifth-grade reading level) to 66 (10th-grade reading level). The VHI, VHI-10, VPQ, VoiSS, and VOS all met the fifth-grade reading level recommended by health literacy experts. In contrast, the VAPP had a 10th-grade reading level.

This last point highlights that perceived deficiencies in initial development are often evaluated in subsequent studies. Many of these PRO measures are used worldwide and have undergone post hoc reliability and validity testing in several languages. Some are now staples in the battery of tests used to screen for, assess, and track improvement in voice-related conditions. Nonetheless, end users must be careful to consider the underlying measurement characteristics of these measures before applying them for a particular purpose. A poorly selected outcome measure can distort clinical or research data, leading to spurious conclusions. In addition, end users must be able to access and personally assess a PRO measure and its items prior to implementation. Nearly all published articles describing the development and validation process (31/32) either listed or provided a link to view the specific items incorporated into the respective PRO measure.

Limitations

Despite the careful design, the search may not have captured all available literature, as poorly indexed literature is often difficult to identify. Hand searches were used to mitigate this limitation. We also limited our search to voice-related PRO measures published in English; those published in other languages were not captured. There is also the risk of subjectivity in the scoring of PRO measure characteristics. Every effort was made to minimize this risk by using three independent reviewers for each instrument considered. Moreover, not all psychometric and applicability parameters were evaluated (e.g., translatability, alternate form reliability [comparability of paper- and Internet-based administrations]). Reviewers were not blinded to the authors of the PRO measures they were evaluating. This could potentially introduce bias; however, blinding of authors is not common practice either by the Cochrane Collaboration or in Agency for Healthcare Research and Quality systematic reviews, which are considered the standards of excellence in systematic review methodology. Last, it is incumbent on the end user to understand the measurement and applicability characteristics of any candidate measure and to select a measure developed for a specified targeted construct in a similar population that corresponds to their specific clinical or research needs and aims.

Clinical Implications

This systematic review has important clinical and research implications. First, it highlights that voice-related PRO measures are varied in psychometric properties. This is critically important because a clinician or researcher considering using a particular instrument should be aware of whether it has been developed and validated properly for the proposed application. For example, a measure that lacks demonstrable responsiveness to change should not be used to track treatment outcomes. The glossary shown in Table 2 describes those parameters that clinicians and researchers should consider when deciding which PRO measure to implement into their practice or research. A poorly designed PRO measure could simply give the wrong answers. Clinicians and researchers rely on these tools to guide care, and they should be able to have faith that they are measuring what they think they are. Many instruments exist, and this systematic review provides some guidance regarding the respective strengths and limitations of available voice-related measures. This review also highlights methodological deficiencies in current measures that may be addressable in future studies.

Conclusions

PRO measures are currently the principal means of evaluating treatment effectiveness in voice-related conditions. Despite their prominence, available PRO measures have disparate psychometric rigor. Two important thematic deficiencies in current voice-related PRO measures are the lack of patient involvement in the item development process and lack of strong construct validity. Issues related to response burden and scoring and interpretation were also highlighted. Care must be taken to understand the measurement properties and utility of PRO measures before selecting and advocating their use for either research or clinical applications.

Acknowledgment

Funding was provided by National Institute on Deafness and Other Communication Disorders Grants K23DC013559 and L30DC012687 to the first author.

References

Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11,* 193–205.

Ahmed, S., Berzon, R. A., Revicki, D. A., Lenderking, W. R., Moinpour, C. M., Basch, E., . . . International Society for Quality of Life Research. (2012). The use of patient-reported outcomes (PRO) within comparative effectiveness research: Implications for clinical practice and health care policy. *Medical Care, 50,* 1060–1070. doi:10.1097/MLR.0b013e318268aaff

Anastasi, A. (1988). *Psychological testing*. New York, NY: Macmillan.

Andrae, D. A., Patrick, D. L., Drossman, D. A., & Covington, P. S. (2013). Evaluation of the Irritable Bowel Syndrome Quality of Life (IBS-QOL) questionnaire in diarrheal-predominant irritable bowel syndrome patients. *Health and Quality of Life Outcomes, 11,* 208. doi:10.1186/1477-7525-11-208

Arffa, R. E., Krishna, P., Gartner-Schmidt, J., & Rosen, C. A. (2012). Normative values for the Voice Handicap Index-10. *Journal of Voice, 26,* 462–465. doi:10.1016/j.jvoice.2011.04.006

Bach, K. K., Belafsky, P. C., Wasylik, K., Postma, G. N., & Koufman, J. A. (2005). Validity and reliability of the Glottal Function Index. *Archives of Otolaryngology—Head & Neck Surgery, 131,* 961–964. doi:10.1001/archotol.131.11.961

Baylor, C., McAuliffe, M. J., Hughes, L. E., Yorkston, K., Anderson, T., Kim, J., & Amtmann, D. (2014). A differential item functioning (DIF) analysis of the Communicative Participation Item Bank (CPIB): Comparing individuals with Parkinson's disease from the United States and New Zealand. *Journal of Speech, Language, and Hearing Research, 57,* 90–95. doi:10.1044/1092-4388(2013/12-0414)

Baylor, C., Yorkston, K., Eadie, T., Kim, J., Chung, H., & Amtmann, D. (2013). The Communicative Participation Item Bank (CPIB): Item bank calibration and development of a disorder-generic short form. *Journal of Speech, Language, and Hearing Research, 56,* 1190–1208. doi:10.1044/1092-4388(2012/12-0140)

Baylor, C. R., Yorkston, K. M., Eadie, T. L., Miller, R. M., & Amtmann, D. (2009). Developing the Communicative Participation Item Bank: Rasch analysis results from a spasmodic dysphonia sample. *Journal of Speech, Language, and Hearing Research, 52,* 1302–1320. doi:10.1044/1092-4388(2009/07-0275)

Belafsky, P. C., Postma, G. N., & Koufman, J. A. (2002). Validity and reliability of the Reflux Symptom Index (RSI). *Journal of Voice, 16,* 274–277.

Benninger, M. S., Ahuja, A. S., Gardner, G., & Grywalski, C. (1998). Assessing outcomes for dysphonic patients. *Journal of Voice, 12,* 540–550.

Biddle, A. K., Watson, L. R., Hooper, C. R., Lohr, K. N., & Sutton, S. F. (2002). Criteria for determining disability in speech-language disorders (Summary, Evidence Report/Technology Assessment: Number 52). Retrieved from Agency for Healthcare Research and Quality (AHRQ) Evidence Report Summaries, https://www.ncbi.nlm.nih.gov/books/NBK11866/

Bogaardt, H. C., Hakkesteegt, M. M., Grolman, W., & Lindeboom, R. (2007). Validation of the Voice Handicap Index using Rasch analysis. *Journal of Voice, 21,* 337–344. doi:10.1016/j.jvoice.2005.09.007

Branski, R. C., Cukier-Blaj, S., Pusic, A., Cano, S. J., Klassen, A., Mener, D., . . . Kraus, D. H. (2010). Measuring quality of life in dysphonic patients: A systematic review of content development in patient-reported outcomes measures. *Journal of Voice, 24,* 193–198. doi:10.1016/j.jvoice.2008.05.006

Carding, P. N., & Horsley, I. A. (1992). An evaluation study of voice therapy in non-organic dysphonia. *European Journal of Disorders of Communication, 27,* 137–158.

Carding, P. N., Horsley, I. A., & Docherty, G. J. (1998). The effectiveness of voice therapy for patients with non-organic dysphonia. *Clinical Otolaryngology and Allied Sciences, 23,* 310–318.

Cheng, J., & Woo, P. (2010). Correlation between the Voice Handicap Index and voice laboratory measurements after phonosurgery. *Ear, Nose and Throat Journal, 89,* 183–188.

Cohen, S. M., Kim, J., Roy, N., Asche, C., & Courey, M. (2012a). Direct health care costs of laryngeal diseases and disorders. *Laryngoscope, 122,* 1582–1588. doi:10.1002/lary.23189

Cohen, S. M., Kim, J., Roy, N., Asche, C., & Courey, M. (2012b). Prevalence and causes of dysphonia in a large treatment-seeking population. *Laryngoscope, 122,* 343–348. doi:10.1002/lary.22426

Deary, I. J., Wilson, J. A., Carding, P. N., & MacKenzie, K. (2003). VoiSS: A patient-derived voice symptom scale. *Journal of Psychosomatic Research, 54,* 483–489.

Dew, K., Keefe, V., & Small, K. (2005). "Choosing" to work when sick: Workplace presenteeism. *Social Science and Medicine, 60,* 2273–2282. doi:10.1016/j.socscimed.2004.10.022

Eadie, T. L., Lamvik, K., Baylor, C. R., Yorkston, K. M., Kim, J., & Amtmann, D. (2014). Communicative participation and quality of life in head and neck cancer. *Annals of Otology, Rhinology & Laryngology, 123,* 257–264. doi:10.1177/0003489414525020

Epstein, R., Hirani, S. P., Stygall, J., & Newman, S. P. (2009). How do individuals cope with voice disorders? Introducing the Voice Disability Coping Questionnaire. *Journal of Voice, 23,* 209–217. doi:10.1016/j.jvoice.2007.09.001

Epstein, R., Stygall, J., & Newman, S. (1997). The short-term impact of Botox injections on speech disability in adductor spasmodic dysphonia. *Disability and Rehabilitation, 19,* 20–25.

Feeny, D. H., Eckstrom, E., Whitlock, E. P., & Perdue, L. A. (2013). *A primer for systematic reviewers on the measurement of functional status and health-related quality of life in older adults* (AHRQ Publication No. 13-EHC128-EF). Rockville, MD: Agency for Healthcare Research and Quality.

Feeny, D., Furlong, W., Boyle, M., & Torrance, G. W. (1995). Multi-attribute health status classification systems. Health Utilities Index. *Pharmacoeconomics, 7,* 490–502.

Feinstein, A. R. (1983). An additional basic science for clinical medicine: IV. The development of clinimetrics. *Annals of Internal Medicine, 99,* 843–848.

Francis, D. O., McPheeters, M. L., Noud, M., Penson, D. F., & Feurer, I. D. (2016). Checklist to operationalize measurement characteristics of patient-reported outcome measures. *Systematic Reviews, 5,* 129.

Franic, D. M., Bramlett, R. E., & Bothe, A. C. (2005). Psychometric evaluation of disease specific quality of life instruments in voice disorders. *Journal of Voice, 19,* 300–315. doi:10.1016/j.jvoice.2004.03.003

Garrow, A. P., Khan, N., Tyson, S., Vestbo, J., Singh, D., & Yorke, J. (2015). The development and first validation of the Manchester Early Morning Symptoms Index (MEMSI) for patients with COPD. *Thorax, 70,* 757–763. doi:10.1136/thoraxjnl-2014-206410

Ghirardi, A. C., Ferreira, L. P., Giannini, S. P., & Latorre Mdo, R. (2013). Screening Index for Voice Disorder (SIVD): Development and validation. *Journal of Voice, 27,* 195–200. doi:10.1016/j.jvoice.2012.11.004

Gillespie, A. I., & Abbott, K. V. (2011). The influence of clinical terminology on self-efficacy for voice. *Logopedics, Phoniatrics, Vocology, 36,* 91–99. doi:10.3109/14015439.2010.539259

Gliklich, R. E., Glovsky, R. M., & Montgomery, W. W. (1999). Validation of a voice outcome survey for unilateral vocal cord paralysis. *Otolaryngology—Head & Neck Surgery, 120,* 153–158.

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., Norman, G. R., & Clinical Significance Consensus Meeting Group. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77,* 371–383. doi:10.1016/S0025-6196(11)61793-X

Guyatt, G., & Schunemann, H. (2007). How can quality of life researchers make their work more useful to health workers and their patients? *Quality of Life Research, 16,* 1097–1105. doi:10.1007/s11136-007-9223-3

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer.

Hill, D. S., Akhtar, S., Corroll, A., & Croft, C. B. (2000). Quality of life issues in recurrent respiratory papillomatosis. *Clinical Otolaryngology and Allied Sciences, 25,* 153–160.

Hirano, M. (1981). *Clinical examination of voice*. New York, NY: Springer-Verlag.

Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice, 13,* 557–569.

Hopkins, C., Yousaf, U., & Pedersen, M. (2006). Acid reflux treatment for hoarseness (Art. No. CD005054). *Cochrane*

Database of Systematic Reviews. doi:10.1002/14651858.CD005054.pub2

Hsiung, M. W., Pai, L., & Wang, H. W. (2002). Correlation between Voice Handicap Index and voice laboratory measurements in dysphonic patients. *European Archives of Oto-Rhino-Laryngology, 259,* 97–99.

Hu, A., Isetti, D., Hillel, A. D., Waugh, P., Comstock, B., & Meyer, T. K. (2013). Disease-specific self-efficacy in spasmodic dysphonia patients. *Otolaryngology—Head & Neck Surgery, 148,* 450–455. doi:10.1177/0194599812472319

Hutchings, H. A., Cheung, W. Y., Russell, I. T., Durai, D., Alrubaiy, L., & Williams, J. G. (2015). Psychometric development of the Gastrointestinal Symptom Rating Questionnaire (GSRQ) demonstrated good validity. *Journal of Clinical Epidemiology, 68,* 1176–1183. doi:10.1016/j.jclinepi.2015.03.019

Isetti, D., & Meyer, T. (2014). Workplace productivity and voice disorders: A cognitive interviewing study on presenteeism in individuals with spasmodic dysphonia. *Journal of Voice, 28,* 700–710. doi:10.1016/j.jvoice.2014.03.017

Jacobson, B., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., & Benninger, M. S. (1997). The Voice Handicap Index (VHI): Development and validation. *American Journal of Speech-Language Pathology, 6,* 66–70.

Johnston, B. C., Ebrahim, S., Carrasco-Labra, A., Furukawa, T. A., Patrick, D. L., Crawford, M. W., . . . Nesrallah, G. (2015). Minimally important difference estimates and methods: A protocol. *BMJ Open, 5*(10), e007953. doi:10.1136/bmjopen-2015-007953

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18,* 124–132. doi:10.1044/1058-0360(2008/08-0017)

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel*. Springfield, VA: National Technical Information Service.

Konnai, R. M., Jayaram, M., & Scherer, R. C. (2010). Development and validation of a voice disorder outcome profile for an Indian population. *Journal of Voice, 24,* 206–220. doi:10.1016/j.jvoice.2008.06.006

Lam, P. K., Chan, K. M., Ho, W. K., Kwong, E., Yiu, E. M., & Wei, W. I. (2006). Cross-cultural adaptation and validation of the Chinese Voice Handicap Index-10. *Laryngoscope, 116,* 1192–1198. doi:10.1097/01.mlg.0000224539.41003.93

Laukkanen, A. M., Leppanen, K., & Ilomaki, I. (2009). Self-evaluation of voice as a treatment outcome measure. *Folia Phoniatrica et Logopaedica, 61,* 57–65. doi:10.1159/000201000

Lehto, L., Rantala, L., Vilkman, E., Alku, P., & Backstrom, T. (2003). Experiences of a short vocal training course for call-centre customer service advisors. *Folia Phoniatrica et Logopaedica, 55,* 163–176.

Llewellyn-Thomas, H. A., Sutherland, H. J., Hogg, S. A., Ciampi, A., Harwood, A. R., Keane, T. J., . . . Boyd, N. F. (1984). Linear analogue self-assessment of voice quality in laryngeal cancer. *Journal of Chronic Diseases, 37,* 917–924.

Ma, E. P., & Yiu, E. M. (2001). Voice activity and participation profile: Assessing the impact of voice disorders on daily activities. *Journal of Speech, Language, and Hearing Research, 44,* 511–524.

Ma, E. P., & Yiu, E. M. (2007). Scaling voice activity limitation and participation restriction in dysphonic individuals. *Folia Phoniatrica et Logopaedica, 59,* 74–82. doi:10.1159/000098340

Mathieson, L. (1993). Vocal tract discomfort in hyperfunctional dysphonia. *Voice, 2,* 40–48.

McHorney, C. A., Ware, J. E., Jr., & Raczek, A. E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31,* 247–263.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549. doi:10.1007/s11136-010-9606-8

Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., . . . de Vet, H. C. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*(3), 313–333. doi:10.1007/s11136-009-9451-9

Monk, M. (1981). Blood pressure awareness and psychological well-being in the health and nutrition examination survey. *Clinical and Investigative Medicine, 4,* 183–189.

Nam, I. C., Bae, J. S., Shim, M. R., Hwang, Y. S., Kim, M. S., & Sun, D. I. (2012). The importance of preoperative laryngeal examination before thyroidectomy and the usefulness of a voice questionnaire in screening. *World Journal of Surgery, 36,* 303–309. doi:10.1007/s00268-011-1347-5

Nanjundeswaran, C., Jacobson, B. H., Gartner-Schmidt, J., & Verdolini Abbott, K. (2015). Vocal Fatigue Index (VFI): Development and validation. *Journal of Voice, 29,* 433–440. doi:10.1016/j.jvoice.2014.09.012

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18,* 301–319.

Paolillo, N. P., & Pantaleo, G. (2015). Development and validation of the Voice Fatigue Handicap Questionnaire (VFHQ): Clinical, psychometric, and psychosocial facets. *Journal of Voice, 29,* 91–100. doi:10.1016/j.jvoice.2014.05.010

Patient-Centered Outcomes Research Institute. (2014). *Working definition of patient-centered outcomes research*. Retrieved from http://www.pcori.org/images/PCOR_Rationale.pdf

Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P., Dawisha, S., . . . Kennedy, D. L. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health, 10*(Suppl. 2), S125–S137. doi:10.1111/j.1524-4733.2007.00275.x

Penson, D. F., Litwin, M. S., & Aaronson, N. K. (2003). Health related quality of life in men with prostate cancer. *Journal of Urology, 169,* 1653–1661. doi:10.1097/01.ju.0000061964.49961.55

Pleil, A. M., Coyne, K. S., Reese, P. R., Jumadilova, Z., Rovner, E. S., & Kelleher, C. J. (2005). The validation of patient-rated global assessments of treatment benefit, satisfaction, and willingness to continue—The BSW. *Value Health, 8*(Suppl. 1), S25–S34. doi:10.1111/j.1524-4733.2005.00069.x

Pouwer, F., Snoek, F. J., van der Ploeg, H. M., Ader, H. J., & Heine, R. J. (2000). The well-being questionnaire: Evidence for a three-factor structure with 12 items (W-BQ12). *Psychological Medicine, 30,* 455–462.

Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., . . . Butt, Z. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research, 22,* 1889–1905. doi:10.1007/s11136-012-0344-y

Regnault, A., Hamel, J. F., & Patrick, D. L. (2015). Pooling of cross-cultural PRO data in multinational clinical trials: How much can poor measurement affect statistical power? *Quality of Life Research, 24*(2), 273–277. doi:10.1007/s11136-014-0765-x

Riehm, K. E., Kwakkenbos, L., Carrier, M. E., Bartlett, S. J., Malcarne, V. L., Mouthon, L., . . . Investigators, Scleroderma Patient-Centered Intervention Network. (2016). Validation of the Self-Efficacy for Managing Chronic Disease (SEMCD) scale: A Scleroderma Patient-Centered Intervention Network (SPIN) cohort study. *Arthritis Care and Research, 68,* 1195–1200. doi:10.1002/acr.22807

Rosen, C. A., Lee, A. S., Osborne, J., Zullo, T., & Murry, T. (2004). Development and validation of the Voice Handicap Index-10. *Laryngoscope, 114,* 1549–1556. doi:10.1097/00005537-200409000-00009

Rosen, C. A., Murry, T., Zinn, A., Zullo, T., & Sonbolian, M. (2000). Voice Handicap Index change following treatment of voice disorders. *Journal of Voice, 14,* 619–623.

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology, 22,* 212–226. doi:10.1044/1058-0360(2012/12-0014)

Roy, N., Merrill, R. M., Gray, S. D., & Smith, E. M. (2005). Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *Laryngoscope, 115,* 1988–1995. doi:10.1097/01.mlg.0000179174.32345.41

Ruotsalainen, J. H., Sellman, J., Lehto, L., Jauhiainen, M., & Verbeek, J. H. (2007a). Interventions for preventing voice disorders in adults (Art. No. CD006372). *Cochrane Database of Systematic Reviews*. doi:10.1002/14651858.CD006372.pub2

Ruotsalainen, J. H., Sellman, J., Lehto, L., Jauhiainen, M., & Verbeek, J. H. (2007b). Interventions for treating functional dysphonia in adults (Art. No. CD006373). *Cochrane Database of Systematic Reviews*. doi:10.1002/14651858.CD006373.pub2

Scott, S., Robinson, K., Wilson, J. A., & Mackenzie, K. (1997). Patient-reported problems associated with dysphonia. *Clinical Otolaryngology and Allied Sciences, 22,* 37–40.

Sherbourne, C. D., & Stewart, A. L. (1991). The MOS Social Support survey. *Social Science and Medicine, 32,* 705–714.

Smith, E., Verdolini, K., Gray, S., Nichols, S., Lemke, J., Barkmeier, J., . . . Hoffman, H. (1996). Effect of voice disorders on quality of life. *Journal of Medical Speech-Language Pathology, 4,* 223–244.

Snyder, C. F., Jensen, R. E., Segal, J. B., & Wu, A. W. (2013). Patient-reported outcomes (PROs): Putting the patient perspective in patient-centered outcomes research. *Medical Care, 51*(8 Suppl. 3), S73–S79. doi:10.1097/MLR.0b013e31829b1d84

Speyer, R., Wieneke, G. H., & Dejonckere, P. H. (2004). Self-assessment of voice therapy for chronic dysphonia. *Clinical Otolaryngology and Allied Sciences, 29,* 66–74.

Teixeira, L. C., Rodrigues, A. L., Silva, A. F., Azevedo, R., Gama, A. C., & Behlau, M. (2013). The use of the URICA-Voice questionnaire to identify the stages of adherence to voice treatment. *Codas, 25,* 8–15.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., . . . de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60,* 34–42. doi:10.1016/j.jclinepi.2006.03.012

Torrance, G. W., Furlong, W., Feeny, D., & Boyle, M. (1995). Multi-attribute preference functions. Health Utilities Index. *Pharmacoeconomics, 7,* 503–520.

van Gogh, C. D., Verdonck-de Leeuw, I. M., Boon-Kamma, B. A., Langendijk, J. A., Kuik, D. J., & Mahieu, H. F. (2005). A screening questionnaire for voice problems after treatment of early glottic cancer. *International Journal of Radiation Oncology, Biology, Physics, 62,* 700–705. doi:10.1016/j.ijrobp.2004.10.027

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231–251.

Verdonck-de Leeuw, I. M., Keus, R. B., Hilgers, F. J., Koopmans-van Beinum, F. J., Greven, A. J., de Jong, J. M., . . . Bartelink, H. (1999). Consequences of voice impairment in daily life for patients following radiotherapy for early glottic cancer: Voice quality, vocal function, and vocal performance. *International Journal of Radiation Oncology, Biology, Physics, 44,* 1071–1078.

Vitali, C., Borghi, E., Napoletano, A., Polini, F., Caronni, M., Ammenti, P., & Cattaneo, D. (2010). Oropharyngolaryngeal disorders in scleroderma: Development and validation of the SLS scale. *Dysphagia, 25,* 127–138. doi:10.1007/s00455-009-9235-y

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30,* 473–483.

Wegener, S. T., Castillo, R. C., Heins, S. E., Bradford, A. N., Newell, M. Z., Pollak, A. N., & MacKenzie, E. J. (2014). The development and validation of the Readiness to Engage in Self-Management after Acute Traumatic Injury questionnaire. *Rehabilitation Psychology, 59,* 203–210. doi:10.1037/a0035693

Wheeler, K. M., Collins, S. P., & Sapienza, C. M. (2006). The relationship between VHI scores and specific acoustic measures of mildly disordered voice production. *Journal of Voice, 20,* 308–317. doi:10.1016/j.jvoice.2005.03.006

Wilson, J. A., Webb, A., Carding, P. N., Steen, I. N., MacKenzie, K., & Deary, I. J. (2004). The Voice Symptom Scale (VoiSS) and the Vocal Handicap Index (VHI): A comparison of structure and content. *Clinical Otolaryngology and Allied Sciences, 29,* 169–174. doi:10.1111/j.0307-7772.2004.00775.x

Woisard, V., Bodin, S., Yardeni, E., & Puech, M. (2007). The Voice Handicap Index: Correlation between subjective patient response and quantitative assessment of voice. *Journal of Voice, 21,* 623–631. doi:10.1016/j.jvoice.2006.04.005

World Health Organization. (2001). *The International Classification of Functioning, Disability and Health (ICF)*. Geneva, Switzerland: Author. Retrieved from http://who.int/classifications/icf/en/

Zraick, R. I., & Atcherson, S. R. (2012). Readability of patient-reported outcome questionnaires for use with persons with dysphonia. *Journal of Voice, 26,* 635–641. doi:10.1016/j.jvoice.2011.01.009

Appendix A

MEDLINE Search Strategies: Dysphonia (PubMed Platform)

Search terms		Search results
#1	"voice disorder"[tiab] OR "voice disorders"[tiab] OR "voice disordered"[tiab] OR "voice handicap"[tiab] OR "vocal handicap"[tiab] OR "vocally handicapped"[tiab] OR "voice handicap"[tiab] OR "voice handicaps"[tiab] OR "voice handicapped"[tiab] OR "vocal disorder"[tiab] OR "vocal disorders"[tiab] OR "vocally disordered"[tiab] OR "vocal disability"[tiab] OR "vocal disabilities"[tiab] OR "voice disability"[tiab] OR "voice disabilities"[tiab] OR dysphonia*[tiab] OR dysphonic[tiab] OR aphonia*[tiab] OR aphonic[tiab] OR "Voice Disorders"[MeSH] OR "Vocal Cords/pathology"[MeSH] OR "Vocal Cord Paralysis"[MeSH] OR "vocal cord paralysis"[tiab] OR "vocal fold paralysis"[tiab] OR hoarse[tiab] OR hoarseness[tiab] OR "speech disability"[tiab] OR "speech disabilities"[tiab] OR "voice symptom"[tiab] OR "voice symptoms"[tiab] OR "vocal symptom"[tiab] OR "vocal symptoms" OR "vocal roughness"[tiab] OR "voice activity"[tiab] OR "vocal fatigue"[tiab] OR "resonance disorder"[tiab] OR "resonance disorders"[tiab] OR phonation[tiab]	18,698
#2	"Psychometrics"[MeSH] OR psychometric*[tiab] OR scale*[tiab] OR score*[tiab] OR inventory[tiab] OR inventories[tiab] OR questionnaire*[tiab] OR Questionnaires[MeSH] OR inventories[tiab] OR index[tiab] OR indices[tiab] OR instrument*[tiab] OR outcome measure*[tiab] OR measurement[tiab] OR "Patient Satisfaction"[MeSH] OR "quality of life"[tiab] OR "Qualitative Research"[MeSH] OR validation studies[pt] OR "Treatment Outcome"[MeSH] OR "Disability Evaluation"[MeSH] OR "Health Surveys"[MeSH] OR "Reproducibility of Results"[MeSH] OR "Severity of Illness Index"[MeSH]	2,991,677
#3	#1 AND #2 AND English[lang] AND Humans[MeSH]	3,878
#4	newspaper article[pt] OR letter[pt] OR comment[pt] OR case reports[pt] OR review[pt] OR practice guideline[pt] OR news[pt] OR editorial[pt] OR historical article[pt] OR meta-analysis[pt] OR legal cases[pt] OR jsubsetk	4,930,194
#5	#3 NOT #4	3,116

Note. [tiab] = title/abstract word; [MeSH] = medical subject heading; [pt] = publication type; [lang] = language; jsubsetk = consumer health literature.

Appendix B

Cumulative Index of Nursing and Allied Health Literature Search Strategies: Dysphonia (EbscoHost Platform)

Search terms		Search results
#1	(MH "Voice Disorders") OR "voice disorders" OR "voice disorder" OR "voice disordered" OR "voice handicap" OR "vocal handicap" OR "vocally handicapped" OR "voice handicaps" OR "voice handicapped" OR "vocal disorder" OR "vocal disorders" OR "vocally disordered" OR "vocal disability" OR "vocal disabilities" OR "voice disability" OR "voice disabilities" OR "dysphonia*" OR "dysphonic" OR (MH "Aphonia") OR "aphonia*" OR "aphonic" OR (MH "Vocal Cords/PA") OR (MH "Vocal Cord Paralysis") OR "vocal cord paralysis" OR "vocal fold paralysis" OR (MH "Hoarseness") OR "hoarse" OR "speech disability" OR "speech disabilities" OR "voice symptom" OR "voice symptoms" OR "vocal symptom" OR "vocal symptoms" OR "vocal roughness" OR "voice activity" OR "vocal fatigue" OR "resonance disorder" OR "resonance disorders" OR (MH "Phonation") OR "phonation"	3,946
#2	(MH "Psychometrics") OR "psychometric*" OR (MH "Scales") OR "scale*" OR "score*" OR (MH "Inventories") OR "inventories" OR "inventory" OR (MH "Questionnaires") OR "questionnaire*" OR "index" OR "indices" OR "instrument*" OR (MH "Instrument Validation") OR (MH "Outcome Assessment") OR "outcome measure*" OR (MH "Treatment Outcomes") OR "measurement" OR (MH "Patient Satisfaction") OR (MH "Quality of Life") OR "quality of life" OR (MH "Qualitative Studies") OR (MH "Validation Studies") OR (MH "Disability Evaluation") OR (MH "Surveys") OR (MH "Reproducibility of Results") OR (MH "Severity of Illness Indices")	627,136
#3	#1 AND #2 Limiters - Exclude MEDLINE records; Human; Language: English	142

Note. MH = medical subject heading.

Appendix C

Health and Psychosocial Instrument Search Strategies: Dysphonia (Ovid Platform)

Search terms		Search results
#1	(“voice disorder$” or “voice handicap$” or “vocal$ handicap$” or “vocal disorder$” or dysphoni$ OR aphoni$ or “vocal cord paralysis” or hoarse$ or “vocal disability$” or “speech disability$” or “voice disabilit$” or “vocal fatigue” or “voice symptom$” or “vocal symptom*” or “vocal roughness”).mp.	17

Note. mp = title, acronym, descriptors, measure descriptors, sample descriptors, abstract, source.