AJSLP

Review

Evidence-Based Clinical Voice Assessment: A Systematic Review

Nelson Roy,[a] Julie Barkmeier-Kraemer,[b] Tanya Eadie,[c] M. Preeti Sivasankar,[d] Daryush Mehta,[e] Diane Paul,[f] and Robert Hillman[e,g]

Purpose: To determine what research evidence exists to support the use of voice measures in the clinical assessment of patients with voice disorders.
Method: The American Speech-Language-Hearing Association (ASHA) National Center for Evidence-Based Practice in Communication Disorders staff searched 29 databases for peer-reviewed English-language articles between January 1930 and April 2009 that included key words pertaining to objective and subjective voice measures, voice disorders, and diagnostic accuracy. The identified articles were systematically assessed by an ASHA-appointed committee employing a modification of the critical appraisal of diagnostic evidence rating system.
Results: One hundred articles met the search criteria. The majority of studies investigated acoustic measures (60%) and focused on how well a test method identified the presence or absence of a voice disorder (78%). Only 17 of the 100 articles were judged to contain adequate evidence for the measures studied to be formally considered for inclusion in clinical voice assessment.
Conclusion: Results provide evidence for selected acoustic, laryngeal imaging–based, auditory–perceptual, functional, and aerodynamic measures to be used as effective components in a clinical voice evaluation. However, there is clearly a pressing need for further high-quality research to produce sufficient evidence on which to recommend a comprehensive set of methods for a standard clinical voice evaluation.

Key Words: voice disorders, assessment, technology, diagnostics

In the past 10 years, there has been an increased call for evidence-based practice (EBP) within the field of speech-language pathology (American Speech-Language-Hearing Association [ASHA], 2005; Dollaghan, 2004; Yorkston et al., 2001). EBP has been defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients [by] integrating individual clinical expertise with the best available external clinical evidence from systematic research" (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996, p. 71). ASHA, through its National Center for Evidence-Based Practice in Communication Disorders (N-CEP), and the Academy of Neurologic Communication Disorders and Sciences have played instrumental roles in conducting evidence-based systematic reviews and establishing clinical practice guidelines, when warranted, for treating individuals with communication disorders. Most of the focus of these reviews and guidelines has been on management, with very few EBPs specified for diagnostics or functional assessment (ASHA, 2008; Coelho, Ylvisaker, & Turkstra, 2005; Turkstra, Coelho, & Ylvisaker, 2005).

In standard clinical voice practice, patients are assessed by a multidisciplinary team consisting of, at minimum, an otolaryngologist and a speech-language pathologist (SLP). Although some of the instruments/tools used during the evaluation may be similar across these team members, the purpose of the evaluation might be different. The physician's objective is to render medical diagnoses related to the identification of laryngeal pathology and to determine appropriate management strategies (e.g., surgery, referral for voice treatment, etc.; Schwartz et al., 2009). The American Academy of Otolaryngology–Head and Neck Surgery recommends that, at a minimum, when evaluating a patient with dysphonia, the basic protocol should include a rigorous clinical history, physical examination, and visualization of the larynx via laryngoscopy (Schwartz et al., 2009).

In contrast, the objective of the SLP is to assess acoustic voice production and its underlying physiological function and to determine how the voice disorder affects an

[a]University of Utah, Salt Lake City
[b]University of California–Davis, Sacramento
[c]University of Washington, Seattle
[d]Purdue University, West Lafayette, IN
[e]Massachusetts General Hospital, Boston, MA
[f]American Speech-Language-Hearing Association, Rockville, MD
[g]Harvard Medical School, Boston, MA
Correspondence to Daryush Mehta: daryush.mehta@alum.mit.edu

Editor: Carol Scheffner Hammer
Associate Editor: Rebecca Leonard
Received February 13, 2012
Revision received July 19, 2012
Accepted October 10, 2012
DOI: 10.1044/1058-0360(2012/12-0014)

individual in everyday situations, as well as to determine prognosis for change, provide recommendations for intervention and support, and recommend referrals where appropriate (ASHA, 1998, 2004). SLPs use a variety of subjective and objective approaches to evaluate voice function. In a survey of experienced SLPs, 100% of 53 respondents reported using auditory–perceptual measures during voice evaluations, followed by observations of body posture and movement and the patient's ability to alter voice production (Behrman, 2005). These types of subjective evaluation methods were significantly more likely to be used than assessments such as laryngostroboscopy or objective measures such as acoustics, aerodynamics, and electroglottography. In fact, the development of objective measures of vocal function stems from a longstanding desire in the field to develop measures that provide more reliable insights into vocal function and are less subjective than assessments that rely heavily on perceptual judgments (see, e.g., Hillman, Montgomery, & Zeitels, 1997).

Although the Behrman (2005) survey highlighted procedures that are most frequently used among voice clinicians, a standardized protocol for the functional assessment of voice pathology does not exist in the United States. This is in contrast to the European Laryngological Society's published clinical guidelines (Dejonckere et al., 2001), which do provide a standard assessment protocol but still lack an adequate evidence base for all of the measures included in it. Procedural inconsistencies and a lack of unified evidence limit the validity and reliability of voice assessment approaches in the United States. This, in turn, precludes the ability to make valid comparisons of vocal function measures with existing normative data and restricts comparisons among facilities, patients, and even repeated assessments of the same patient.

Clinicians routinely use assessment measures to diagnose pathology, confirm the presence or absence of a disease or condition, monitor the progression of a disease or function, assess prognosis, and screen populations who are at risk for disorders (Dollaghan, 2007). The accuracy of test measures is critical. When inaccurate, tests can prevent individuals from receiving optimal treatment in a timely manner. Diagnostic outcome measures describe the probabilistic relationships between positive and negative test results with respect to a reference standard (e.g., laryngoscopy in the voice clinic). Test accuracy is only one of many appraisal points that are commonly used to evaluate the importance of diagnostic evidence. Other appraisal points include subject demographics, recruitment procedures, investigator blinding, and overall importance of findings. For further information related to EBP, the reader is referred to several articles that outline approaches for the critical appraisal of diagnostic evidence (Bossuyt et al., 2003; Dollaghan, 2007; Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000).

The ASHA Special Interest Division for Voice and Voice Disorders (SID 3; now Special Interest Group 3) has been involved in several attempts to systematically review voice assessment methods. One major outcome was creation of the *Classification Manual for Voice Disorders—I* (Verdolini, Rosen, & Branski, 2006), which sought to mirror the classification structure of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000). The purpose of these types of reviews was to (a) better define the state-of-the-art in clinical voice assessment, (b) identify voice assessment areas in need of further research to improve clinical utility, and (c) begin developing recommended guidelines and disorder classifications for clinical voice evaluation. These efforts were motivated primarily by the prevailing lack of evidence-based guidelines for clinical voice assessment, which left individual facilities and clinicians to develop their own protocols, which are used with varying consistency.

Initially, SID 3 focused on auditory–perceptual evaluation of voice, which is the most commonly used clinical assessment procedure (Behrman, 2005). The first tangible result of these efforts was development of the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V; Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009), which was created at a consensus conference that was held by SID 3 at the University of Pittsburgh in June 2002. The conference brought together an international group of voice scientists, experts in human auditory perception, and SLPs, with the explicit goal of creating a standardized protocol useful to clinicians and researchers that would incorporate multiple recommendations for best practices in assessing perceived abnormal voice quality. SID 3 subsequently made the CAPE-V instrument available to the speech-language pathology community for review, trial usage, and comments (Kempster et al., 2009), which was followed by more formal clinical testing (Zraick et al., 2011). There is ongoing interest in trying to improve the reliability of the CAPE-V by creating anchor stimuli and providing standard training modules using established methods (Eadie & Baylor, 2006). Other approaches for improving voice evaluation continue to be investigated (Gerratt & Kreiman, 2001).

In late 2007, ASHA's Executive Board (Resolution) established the Ad Hoc Working Group on Clinical Voice Assessment, which was charged with carrying out an expansive systematic review of commonly used voice assessment methods. Members of the working group included Robert Hillman (chair); Nelson Roy, Tanya Eadie, Julie Barkmeier-Kraemer, M. Preeti Sivasankar, and Diane Paul (ex officio). Two working group members (Christine Sapienza and Dimitar Deliyski) contributed to the formation of the clinical questions and systematic review methodology but were unable to continue membership. The working group collaborated with N-CEP to conduct a systematic literature review to determine whether a standard voice assessment protocol could be recommended for routine clinical use based on the available literature.

Systematic reviews of diagnostic test accuracy aim to determine how good a particular test is at detecting a target condition; in this case, a particular voice disorder. In this systematic review, both of the terms *test* and *condition* are interpreted in a broad sense. Test refers to any procedure

that is designed to acquire information on the vocal status of an individual. Procedures can include the case history, external physical examination (e.g., palpating the neck to assess muscle tension), functional measures (self-report questionnaires), auditory–perceptual evaluation, visual–perceptual evaluation and/or automatic processing of endoscopic images, and laboratory tests (primarily acoustics, aerodynamics, and electroglottography). Similarly, condition refers to the vocal status of an individual, including the presence (normophonic vs. dysphonic) and type of disease/disorder (e.g., cancer vs. nodules vs. polyps vs. muscle tension dysphonia vs. spasmodic dysphonia), as well as monitoring the severity of a voice disorder (e.g., mild, moderate, severe).

Studies of diagnostic accuracy also permit calculation of statistics that provide an index of test performance. Test accuracy is expressed most often in terms of *sensitivity* (i.e., the proportion of individuals with the condition—as determined by the reference standard—who are also positive according to the index test) and *specificity* (i.e., the proportion of individuals without the condition who are also negative according to the index test). Many alternative measures are in use, including positive and negative likelihood ratios, positive and negative predictive values, and receiver operating characteristic curves. Although information regarding the diagnostic accuracy of a test is important, the quality of the study on which those estimates are predicated is equally important.

The purpose of this paper is to report the results of the working group's systematic review of the literature, which was designed to critically appraise the research evidence that exists to support the use of voice measures in the clinical assessment of patients with voice disorders. The quality of individual studies in this review was assessed for internal and external validity, the degree to which estimates of diagnostic accuracy were biased, and the degree to which the results of a study could be applied to clinical practice.

Method

Search Strategy and Study Eligibility

N-CEP staff conducted a systematic search of the literature using 29 electronic databases (see Appendix A) and a core set of key words and expanded search terms pertaining to voice disorders and diagnostic accuracy (Appendix B). The search covered articles that had been written in English and published in a peer-reviewed journal between January 1930 and April 2009. The beginning date was chosen to encompass the advent of formal studies of voice, including the earliest use of technologies like high-speed imaging and sound spectrography; in addition, the date range encompasses the start of the first peer-reviewed journals on voice and speech. The end date represents when N-CEP initiated the final version of the review. Several factors delayed completion of this review, including changes in the membership of the working group as well as the time-consuming nature of actually assessing all of the identified literature. N-CEP rules (i.e., systematic reviews are only repeated every 5 years) and the desire to maintain the integrity of the review methodology precluded adding references that have appeared since April 2009. It is anticipated that the systematic review will be updated in 2014.

Studies that were incorporated in the review had to include key words (see Appendix B) pertaining to test measures from one or more of the following eight categories of assessment procedures: case history, auditory–perceptual judgments, aerodynamic measures, functional measures, acoustic measures, imaging measures, physical exams, and electroglottography.

Studies included in the review also had to use one or more of these eight assessment procedures to address one or more of the following three clinical questions:

- What is the evidence that a given assessment procedure(s) is capable of determining the *presence/absence* of a voice disorder?
- What is the evidence that a given assessment procedure(s) is capable of determining the *nature (etiology)* of a voice disorder?
- What is the evidence that a given assessment procedure(s) is capable of determining the *extent (severity)* of a voice disorder?

The search excluded studies if they used animal models or pharmacological interventions or if they consisted of participants with selective mutism or resonance disorders or users of alaryngeal speech (e.g., total laryngectomy, artificial larynx, electrolarynx, pseudoglottis). Also, studies investigating invasive electromyographic measures were not included in this review because of the desire to assess measures that could be obtained by SLPs.

Data Abstraction

Each of the five members of the working group was assigned a subset of the articles to assess the quality of the evidence contained within. For the purpose of this systematic review, raters used a modified version of the Critical Appraisal of Diagnostic Evidence (CADE) form (Dollaghan, 2007). The CADE form was originally developed by synthesizing criteria and questions from several diagnostic reporting evaluation systems used in the field of medicine (Bossuyt et al., 2003; Sackett et al., 2000). The modified CADE form (CADE–M) evaluates assessment measures or procedures designed for the purpose of screening, diagnosis, or differential diagnosis. The critical comparison is between a test measure (i.e., the diagnostic procedure under evaluation) and what is referred to as the reference standard. The *reference standard* is defined typically as the "best available method for establishing presence or absence of the target condition" (Bossuyt et al., 2003, p. 8). The working group piloted the CADE–M form with three publications matching the search criteria. Based on experience gained in this pilot phase, the form was modified and finalized.

Appendix C shows the version of the CADE–M form that we used to assess each article; the form helps to assess the merits of an article on a series of appraisal points. Some

of the appraisal points require a simple binary judgment; other ratings are more subjective. During the subsequent data abstraction process, any existing disagreements in rating methodology were resolved by group consensus.

Reliability of Data Abstraction

In order to evaluate the interrater reliability of data abstraction, 20% of the original articles were randomly selected and blindly rerated by a second reviewer on the committee. The working group calculated exact agreement values as measures of reliability for evaluation parameters. In addition, a sensitivity analysis determined whether any disagreements resulted in changes to the overall assessment of a study's quality. Among the articles rated for reliability, a third rater reconciled any discrepancies to determine final data entries for the purpose of assessing the article's evidence-based quality.

Results for interrater reliability of data abstraction are presented separately for the items on the CADE–M that provided basic information about the study (Appendix C, Items 1–6) and for the appraisal points (Appendix C, Appraisal Points A–J). Appraisal Point K was not subject to reliability analysis because responses were not categorical in nature, reflecting the nuance required to recommend or reject a particular article.

Reliability of study information. Categorization of the clinical question (Item 3) ranged in exact agreement from 80% to 90%. Abstraction of participant characteristics (e.g., sample sizes; Item 4) also resulted in exact agreement, ranging from 80% to 90%. Ratings of whether endoscopic visualization was used to verify normality resulted in 90% agreement (Item 4). The categorization of specific assessment measures (Item 5) resulted in some variability across reviewers, with lower agreement for some judgments (e.g., 65% agreement for the auditory–perceptual category) to perfect agreement for others (e.g., 100% agreement for the aerodynamic category). The average interrater agreement across assessment measures was 91%.

Reliability of appraisal points. Agreement ranged from 45% to 100% for Appraisal Points A–J, with a mean agreement of 68% across the appraisal points. In most cases, errors were results of typography, differences of opinion as to whether likelihood ratios could be calculated, or misinterpretations of study design and subject sample sizes.

The final results were found to be 95% accurate with regard to the three criteria (diagnostic accuracy, validity of findings, importance of findings [Appraisal Points G, I, and J, respectively]) that the group used to stratify articles according to evidence-based quality (Levels 1–5). That is, for each criterion, one out of every 20 articles was at risk for being misidentified. This level of agreement was deemed acceptable for the review.

Assessment of Evidence-Based Quality

The results from the CADE-based review were used to determine whether each article met one of five levels of evidence-based quality in terms of diagnostic accuracy and research design. Diagnostic accuracy was judged as being *adequate* or *not adequate*. Adequate diagnostic accuracy required both a positive likelihood ratio >10 and a negative likelihood ratio <0.10 (Appendix C, Appraisal Point G). In some articles, the likelihood ratio statistic was not calculable, which weakened the study's impact.

The quality of each study's research design was based on ratings for the appraisal points that assessed the validity of findings and the importance of findings (Appendix C, Appraisal Points I and J, respectively), which are designed to be integrated judgments over the other appraisal points, as explained on the CADE–M form in Appendix C. Each of these items was rated on a 3-point scale (*compelling – suggestive – equivocal*) to reflect an overall assessment of the research design and methods (validity of findings) and the potential of the results to significantly influence clinical methodologies (importance of findings; Dollaghan, 2007).

The five levels of evidence-based quality are each defined below. The three criteria for determining quality (i.e., diagnostic accuracy, validity of findings, importance of findings) were progressively relaxed from the highest (1) to the lowest (5) level in an attempt to capture all studies that might contribute significant findings to the clinical questions of interest.

- **Level 1:** A*dequate* diagnostic accuracy, *compelling* validity of findings, and *compelling* importance of findings. As outlined by Dollaghan (2007), if both validity and importance of findings are rated as compelling, there is serious evidence to consider adopting the test measure for diagnostic decision making.
- **Level 2:** *Compelling* validity of findings and *compelling* importance of findings, *regardless of* the adequacy of diagnostic accuracy.
- **Level 3:** *Compelling* importance of findings, *regardless of* ratings for diagnostic accuracy and validity of findings.
- **Level 4:** *Compelling* validity of findings, *regardless of* ratings for diagnostic accuracy and importance of findings. As indicated by Dollaghan (2007), even if validity is equivocal in the presence of compelling importance of findings (or vice versa), a change might be considered based on particular patient characteristics.
- **Level 5:** *Adequate* diagnostic accuracy, *suggestive* validity of findings, and *suggestive* importance of findings. Dollaghan (2007) suggests that, in this case, clinicians who base their decisions on the evidence might reach different decisions about whether to use the given test measure.

Results

Study Selection

The initial systematic search of the 29 electronic databases (see Appendix A) yielded 1,077 studies. Two reviewers independently assessed all citations (with 89% agreement) for applicability based on the established

inclusion and exclusion criteria. Any disagreements between the two reviewers were brought to the working group and were resolved through consensus. These same two reviewers also examined the reference lists from all relevant articles to identify additional citations for possible inclusion. A total of 977 studies failed to meet the inclusion criteria and were excluded from further review.

One hundred articles were identified that met the inclusion and exclusion criteria for this review. Note that, after performing the systematic review, the working group recognized that the imaging category consisted of two subcategories—namely, visual–perceptual judgments and image processing measures—that are defined analogously with the categories of auditory–perceptual judgments and acoustic measures, respectively. Table 1 displays the classification of reviewed articles by test measure category and clinical question addressed. The majority of the articles (78%) addressed the clinical question regarding the ability of a test measure to determine the presence or absence of a voice disorder. The ability of a test measure to determine the nature of a voice disorder was addressed in 44% of the articles. The least number of articles (18%) was dedicated to investigating the ability of a given test measure to determine the severity of a voice disorder.

The most frequently studied test measure category was acoustic measures (60%), followed by measures from image processing (32%) and auditory–perceptual judgments (30%). Electroglottography (11%) and aerodynamic measures (10%) were less frequently studied. Five or fewer studies were dedicated to investigations of case history, visual–perceptual judgments of endoscopic images, functional measures, and physical exams. Across all test measures of interest, the quality of each study varied widely.

Evidence-Based Quality

A total of 17 out of the 100 articles reviewed met the criteria for one or more of the five levels of evidence-based quality. Table 2 provides a summary of these 17 articles categorized by clinical question addressed, test measure evaluated, and level of evidence-based quality (Levels 1–5).

Only four articles met the optimal criteria for Level 1 of evidence-based quality (Articles 1–4). All four articles only addressed the clinical question regarding the ability of a test measure to determine the presence or absence of a voice disorder. Three studies focused on acoustic measures (Articles 1–3); the fourth study compared case history information, auditory–perceptual judgments (CAPE-V), visual–perceptual judgments of laryngeal endoscopic images, and functional measures (e.g., voice handicap index; VHI; Jacobson et al., 1997) to determine that functional measures were most sensitive in detecting postthyroidectomy dysphonia (as verified by endoscopic laryngeal examination), followed closely by auditory–perceptual measures (Article 4).

Two additional articles met the criteria for Level 2 of evidence-based quality (Articles 5 and 6). Article 5 demonstrated that visual–perceptual judgments of endoscopic images can be accurately used by gastroenterologists to determine the presence or absence of laryngeal pathology, and Article 6 showed that auditory–perceptual measures alone are not accurate metrics for determining the presence/absence or nature (the target disorder was vocal nodules in children) of a voice disorder.

Another two articles met the criteria for Level 3 of evidenced-based quality (Articles 7 and 8). Article 7 investigated the ability to identify laryngeal cancer (presence/absence) using image processing techniques, and Article 8 demonstrated that a combination of auditory–perceptual and acoustic measures yielded better accuracy than each taken separately in determining the presence/absence or extent (severity) of a voice disorder (Article 8).

Level 4 of evidence-based quality was achieved by three additional publications (Articles 9–11). Article 9 demonstrated that aerodynamic measures could accurately determine the presence/absence of a voice disorder but were not as successful at determining the nature of the disorder. Article 10 concluded that the acoustic measure of vocal range profile was able to determine the presence of a voice disorder in Cantonese women. Article 11 compared the accuracy of acoustic and auditory–perceptual measures in identifying the presence/absence of laryngeal carcinoma in postradiation

Table 1. The number of studies classified by test measure category and clinical question.

Test measure category	Presence/absence of voice disorder	Nature (etiology) of voice disorder	Extent (severity) of voice disorder	Total unique articles[a]
Case history	3	4	0	5
Auditory–perceptual judgments	21	8	13	30
Visual–perceptual judgments	3	3	0	5
Aerodynamic measures	6	4	7	10
Functional measures	4	3	0	5
Acoustic measures	50	17	18	60
Image processing measures	24	15	2	32
Physical exams	1	1	0	1
Electroglottography	9	5	2	11
Total unique articles[a]	78	44	18	

[a]Each article is counted only once within a given category.

Table 2. Summary of articles that met the criteria for one or more levels of evidence-based quality.

Article title	Article author(s)	Publication year	Clinical question	Test measure	Evidence-based quality
1. Laryngeal pathology detection by means of class-specific neural maps	Hadjitodorov, S., Boyanov, B., & Teston, B.	2000	P/A	AC	1–4
2. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection	Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., & Moroz, I. M.	2007	P/A	AC	1–4
3. Identification of pathological voices using glottal noise measures	Parsa, V., & Jamieson, D. G.	2000	P/A	AC	1–4
4. Prospective trial of voice outcomes after thyroidectomy: Evaluation of patient-reported and clinician-determined voice assessments in identifying postthyroidectomy dysphonia	Stojadinovic, A., Henry, L. R., Howard, R. S., Gurevich-Uvena, J., Makashay, M. J., Coppitt, G. L., … Solomon, N. P.	2008	P/A	CH, A-P, F, V-P	1–4
5. Accuracy of laryngeal examination during upper gastrointestinal endoscopy for premalignancy screening: Prospective study in patients with and without reflux symptoms	Cammarota, G., Galli, J., Agostino, S. l., De Corso, E., Rigante, M., Cianci, R., … Gasbarrini, G.	2006	P/A	V-P	2–4
6. Clinician's accuracy in detecting vocal nodules	Dice, G., & Shearer, W. M.	1973	P/A, N	A-P	2–4
7. Spectrometric measurement in laryngeal cancer	Arens, C., Reussner, D., Neubacher, H., Woenckhaus, J., & Glanz, H.	2006	P/A	I	3
8. Classification of dysphonic voice: Acoustic and auditory–perceptual measures	Eadie, T. L., & Doyle, P. C.	2005	P/A, E	A-P, AC	3
9. Receiver operating characteristic analysis of aerodynamic parameters obtained by airflow interruption: A preliminary report	Jiang, J., & Stern, J.	2004	P/A, N	AD	4
10. Reliability of speaking and maximum voice range measures in screening for dysphonia	Ma, E., Robertson, J., Radford, C., Vagne, S., El-Halabi, R., & Yiu, E.	2007	P/A	AC	4
11. Acoustic voice analysis in different phonetic contexts after larynx radiotherapy for T1 vocal cord carcinoma	Rovirosa, A., Ascaso, C., Abellana, R., Martinez-Celdran, E., Ortega, A., Velasco, M., … Biete, A.	2008	P/A	AC	4
12. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors	Godino-Llorente, J. I., & Gomez-Vilda, P.	2004	P/A	AC	5
13. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech	Parsa, V., & Jamieson, D. G.	2001	P/A	AC	5
14. Diagnostic sensitivity and specificity of laryngoscopic signs of reflux laryngitis	Pribuišienë, R., Uloza, V., & Kupcinskas, L.	2008	P/A N	V-P	5
15. Differentiation of adductor-type spasmodic dysphonia from muscle tension dysphonia by spectral analysis	Rees, C. J., Blalock, P. D., Kemp, S. E., Halum, S. L., & Koufman, J. A.	2007	N	AC	5
16. Discrimination of pathological voices using a time-frequency approach	Umapathy, K., Krishnan, S., Parsa, V., & Jamieson, D. G.	2005	P/A	AC	5
17. Integrating global and local analysis of color, texture and geometrical information for categorizing laryngeal images	Verikas, A., Gelzinis, A., Bacauskiene, M., & Uloza, V.	2006	N	I	5

Note. Clinical Question: P/A = presence/absence of a voice disorder; N = nature (etiology) of a voice disorder; E = extent (severity) of a voice disorder. Test Measures: AC = acoustic measures; CH = case history; A-P = auditory–perceptual judgments; F = functional measures; V-P = visual perceptual judgments of endoscopic images; I = image processing measures; AD = aerodynamic measures. Evidence-Based Quality: 1 = adequate precision, compelling validity, and compelling importance; 2 = compelling validity and compelling importance, regardless of precision; 3 = compelling importance, regardless of precision and validity; 4 = compelling validity, regardless of precision or importance; 5 = adequate precision, suggestive validly, and suggestive importance.

patients. Although diagnostic accuracy was inadequate, a spontaneous speech context was shown to be best for demonstrating differences between normal and abnormal voice production.

Finally, six additional articles fit the criteria for Level 5 of evidence-based quality (Articles 12–17). Of these, three showed some success in using acoustic measures to detect the presence/absence of voice disorders (Articles 12, 13, and 16), whereas a fourth study (Article 15) used spectrographic (acoustic) analysis to accurately differentiate between adductor spasmodic dysphonia and muscle tension dysphonia; that is, to determine the specific nature (etiology) of the

voice disorder. The two remaining Level 5 studies showed adequate precision for assessing endoscopic images using either visual–perceptual judgments to determine the presence/absence and nature of voice disorders (Article 14) or digital processing to automatically classify images into broad categories including healthy, diffuse pathology, or nodular pathology, thus indicating the general nature of a voice disorder (Article 17).

Discussion

The results of this systematic review are discussed with respect to the three main goals that are being pursued by the ASHA SIG 3 as cited in the introduction of this report: (a) better define the state-of-the-art in clinical voice assessment, (b) identify voice assessment areas in need of further research to improve clinical utility, and (c) begin developing recommended guidelines and disorder classifications for clinical voice evaluation.

State-of-the-Art in Clinical Voice Assessment

Based on the current review of 100 articles that met the inclusion and exclusion criteria, the majority of studies investigated acoustic measures, with the next most frequently studied methods being laryngeal imaging–based assessments and auditory–perceptual judgments. Interestingly, the focus of most studies was to determine how well the test method identified the presence or absence of a voice disorder. The second most frequent focus was to determine the nature, or etiology, of voice disorders. Few publications addressed the use of test methods to characterize the extent, or severity, of voice disorders.

Only 17 of the 100 articles that we reviewed met one or more of the five levels of evidence-based quality that were employed, thus demonstrating some evidence of the kind of rigor in research design, analysis, interpretation, and consideration of diagnostic accuracy and clinical importance that was required. All 17 of these articles contained adequate evidence for the methods studied to be considered for inclusion in (in cases of positive results), or exclusion from (in cases of negative results), clinical voice assessment. The CADE–M system (Dollaghan, 2007) used in this study included a final appraisal point—"Clinical Bottom Line" (see Appendix C)—that asked judges to make a global decision about whether the overall evidence "does or does not support considering a change in one's current diagnostic approach—whether to adopt a new diagnostic measure or to stop using an inadequate diagnostic measure." In responding to this point, the judges tended not to provide definitive (binary) responses, but rather offered a more in-depth appraisal of the relative strength of the evidence and/or whether additional research/development was needed before a given method could be readily adopted for routine clinical use. Thus, the responses for the final appraisal point could not be easily categorized and included with the rest of the data for the 17 articles shown in Table 2. Instead, the additional information pertaining to the clinical bottom line is briefly summarized as part of the discussion in the following paragraphs.

A majority (8 of 17) of the selected studies (see Table 2) demonstrated that acoustic measures alone could be highly accurate in determining the presence/absence of a voice disorder (Articles 1–3, 10–13, 15, and 16), with another study showing that acoustic spectral analysis can potentially detect the nature of a disorder in the specific circumstance of differentiating vocal hyperfunction from spasmodic dysphonia (Article 15). Additional studies showed that coupling auditory perception with acoustic measures produced higher accuracy in determining the presence/absence and extent of voice disorders (Article 8) than either approach alone, and that auditory perception alone lacked accuracy when applied to determining the presence/absence and specific nature of a voice disorder (Article 6). These 10 studies employed a wide variety of acoustic analysis methods, from traditional signal processing techniques (e.g., spectral, perturbation, signal-to-noise measures) to nonlinear approaches (e.g., fractal measures), with some use of additional algorithms (e.g., neural networks) for optimizing acoustic-based detection schemes.

Although all of the acoustic approaches showed some promise for use in clinical voice assessment, it was determined that most of them would require further development and/or more robust testing before being recommended/employed for routine clinical use. Further development for clinical use would include attempts to automate and/or create efficient user interfaces for selected experimental measures that are not currently implemented in standard software (Articles 1–3, 12, 16). Persistent research design–related issues included biased or limited sample selection (Articles 10, 11), use of nonblinded judges (Article 15), lack of control subjects (Article 15), reliance on retrospective methods (Articles 1–3, 11–13, 15,16), failure to specify disorder severity (Articles 10, 12, 13, 15, 17), and lack of other important study details (Articles 12, 17).

Four of the 17 selected articles described methods for assessing laryngeal endoscopic images. Two of these studies demonstrated that image processing techniques have the potential to accurately detect the presence of a disorder in screening for laryngeal cancer (autofluorescence, Article 7) and to determine the nature of disorders by categorizing laryngeal images as being healthy versus having diffuse or nodular pathology (integration of color, texture, and geometric information, Article 17). Although promising, both image processing methods would benefit from further testing on better defined subject/patient groups. The other two studies demonstrated that visual–perceptual judgments of laryngeal images could be highly accurate in detecting the presence of a disorder in screening for laryngeal disease during routine upper gastrointestinal endoscopy (Article 5), and could also reliably determine the nature of a disorder via the specific laryngoscopic signs that are most indicative of reflux laryngitis (Article 14).

One of the remaining two articles in Table 2 demonstrated that a combination of aerodynamic measures could accurately determine the presence, but not necessarily the

nature, of a voice disorder (Article 9). The second remaining article showed that a functional measure (VHI) was highly accurate in detecting the presence of postthyroidectomy dysphonia (as detected based on videolaryngoscopy) and performed better than case history or auditory–perceptual judgments (Article 5), although auditory–perceptual assessment using the CAPE-V (Kempster et al., 2009) was a close second.

Most studies that included a standard reference for comparison used laryngeal imaging to confirm diagnostic classifications of voice disorders. The consistent use of laryngeal imaging as a standard reference raises the issue as to whether experts already agree that imaging is critical for determining the presence/absence or nature/etiology of a voice disorder. None of the studies directly addressed this consideration.

There were hundreds of additional articles that dealt with the development and testing of voice assessment methods that were excluded from a complete formal review, as evidenced by the 1,077 studies that were initially identified on the basis of containing key words pertaining to one or more of the eight categories of assessment procedures (see Method section). Many of these articles reported promising technical innovations in voice assessment methods and/or results from studies employing procedures that appear to have good potential as clinical voice assessment tools. The biggest factor, however, in limiting the review to 100 articles was the requirement that studies include some measure related to diagnostic accuracy. This very strict criterion was chosen because such information was deemed critical to addressing the basic questions being posed by the review about the diagnostic performance of voice measures. It also admittedly had the secondary benefit of limiting the number of articles reviewed to a more reasonable number and served to clearly demonstrate the need to formally evaluate diagnostic accuracy in future studies of voice assessment methods that are aimed at developing clinical applications.

Overall, the results of the systematic review provided evidence that selected acoustic, laryngeal imaging–based, auditory–perceptual, functional, and aerodynamic measures have the potential to be used as effective components in a clinical voice evaluation, with most of the evidence demonstrating the capability of measures to detect the presence of a voice disorder. The adoption, however, of many of the identified measures into routine clinical practice will require additional development and more robust testing to further strengthen the evidence base, particularly with respect to demonstrating that selected measures are capable of determining the nature and extent of voice disorders, in addition to detecting whether disorders are present.

With respect to objective measures of vocal function (e.g., acoustic, aerodynamic, image processing, etc.), there has been a longstanding interest in developing such methods because they are considered less subjective than assessments that rely heavily on perceptual judgments, thus potentially providing more reliable information/insights concerning vocal function (Hillman et al., 1997). At the very least, the results of the current review provide support for the continued use of objective measures as supplemental diagnostic tools that are used as part of a comprehensive, integrated evaluation to more fully document vocal function and enhance the quality of diagnosis and treatment (Hillman et al., 1997).

Future Research Needs

There is increasing emphasis on EBP in all facets of health care, including speech-language pathology (ASHA, 2005; Dollaghan, 2004; Sackett et al., 1996). The results of the current systematic review point to the need to pursue the type of high-quality clinical research that is required to expand the evidence base for clinical voice assessment methods. Achieving this goal should lead to more accurate, efficient, and cost-effective diagnosis of voice disorders and should provide a better basis for evaluating the efficacy of the methods used to treat voice disorders.

The current review clearly revealed that some test measure categories were studied more frequently than others. For example, acoustic, laryngeal imaging–based, and auditory–perceptual methods were among the most commonly used methods. Interestingly, these measures also are among the most frequently used clinical assessment tools (Behrman, 2005). In contrast, few studies investigated the contribution of case history, aerodynamic measures, functional measures, physical exams, and electroglottography to address one of the clinical questions. The poor representation of the latter methods should not be interpreted to mean that there is no clinical merit in their use. Rather, the paucity of information related to these methods should be a focus for future studies to fill in gaps in the community's knowledge base regarding their contributions to the clinical questions that were focused on in this review.

This systematic review identified many inadequacies in the extant voice literature. One of the most important findings of this review was that the majority of the research articles evaluated offered an inadequate description of methodology, analysis, and results necessary to determine diagnostic accuracy, validity, and clinical importance of findings. These results are consistent with limitations reported in a review of the medical literature (Reid, Lachs, & Feinstein, 1995), and provided the impetus for systematic approaches to reporting diagnostic accuracy (Bossuyt et al., 2003), including those in the field of speech-language pathology (Dollaghan, 2007). Likewise, in this study, reviewers frequently needed to manually derive diagnostic accuracy statistics based on the published results. In many cases, articles might have been judged inadequate for diagnostic accuracy solely due to insufficient information being provided to calculate this important variable. It is strongly recommended that future studies aimed at addressing one of the three clinical questions posed in this systematic review should provide explicit values of sensitivity, specificity, likelihood ratios (positive and negative), and other diagnostic metrics (Dollaghan, 2007).

Another common finding was the lack of scientific rigor put forth in these studies, as judged by the omission or inadequate composition or definition of control groups,

standard references, bias controls, and reliability and validity procedures. Many studies used retrospective research designs without blinding reviewers or determining the reliability of measures or ratings. In addition, studies frequently used participants of mixed etiologies or did not use a standard reference for comparisons with dependent measures. Enforcement of scientific rigor is needed in future research to control for confounding variables that complicate interpretation of clinical research findings.

Consistency in the classification of voice disorders might also be helpful in future research (see, e.g., Verdolini et al., 2006). This recommendation is based on the need to develop a conceptual framework on which hypotheses are formulated and tested. The description of individuals having or not having a voice disorder might be too broad a definition for sufficiently addressing clinically relevant questions. It is more appropriate to specify the diagnoses of interest (such as vocal nodules vs. adductor-type spasmodic dysphonia) than to use generic terms such as "organic lesion" and "neurological pathology." Many acoustic methods aiming to distinguish between types of voice disorders use a variety of broad-to-specific operational definitions for comparing etiologies. Such variations in formulating research questions could affect the diagnostic accuracy of the outcome measures.

An ongoing problem in studies of voice measures is the lack of consensus regarding standards for technical specifications of instrumentation (e.g., frequency response, recording environment, and analysis algorithms) and data acquisition protocols (e.g., subject tasks), which makes it difficult or impossible to compare results across different investigations. Future efforts to assess the diagnostic performance of voice measures would be greatly assisted by first establishing a minimal set of recommended guidelines (perhaps via expert consensus) that is supported by current knowledge about the (a) physiology and acoustics of voice production, (b) etiology and pathophysiology of voice disorders, and (c) technology for assessing vocal function. The guidelines would integrate the work that has already been accomplished in developing the CAPE-V (voice tasks, etc.) and the recommendations of the Workshop on Acoustic Voice Analysis (Titze, 1995), as well as other appropriate sources. The guidelines would include specifications for instrumentation, specification of environmental factors that must be controlled to obtain valid measures, well-defined voice tasks, and well-defined methods for analyzing recorded materials. Technical and protocol details that cannot be standardized could then be isolated as study variables.

Spectrum effects. An optimal study of a diagnostic test includes a broad spectrum of persons who would normally undergo the test in a clinical setting (Dollaghan, 2007). Few studies assessed or considered spectrum effects or bias, possibly leading to inflated estimates of diagnostic accuracy. This broad spectrum is necessary to produce valid, precise, and generalizable estimates of test performance. Test performance often varies, however, with patient characteristics such as age, sex, and severity of disease. For instance, a large number of studies in this review used acoustic analysis techniques to classify normophonic and dysphonic individuals. Rarely was any attention paid to the spectrum of dysphonia severity represented within the disordered group and whether the performance of the acoustic test varied according to the severity of dysphonia. For instance, it is a relatively easy diagnostic task to distinguish severely dysphonic speakers from normophonic speakers. It is a much more difficult acoustic task, however, to reliably distinguish a mildly dysphonic speaker from a vocally normal speaker. Without some effort by researchers to assess the influence of such spectrum effects on diagnostic test performance, difficulties remain in the interpretation of reported results to assess their clinical value (Mulherin & Miller, 2002).

Recommended Guidelines for Clinical Voice Evaluation

Even though this systematic review provided some evidence that selected test measures could be used as effective components in a clinical voice evaluation, it did not produce sufficient evidence on which to recommend a comprehensive set of methods for a standard clinical voice evaluation. The evidence that laryngeal imaging–based measures are able to detect the presence and nature of voice disorders, coupled with the frequent use of laryngoscopy as a reference standard, argues for the continued use of laryngeal endoscopy as a primary diagnostic tool; however, formal studies are still needed to fully validate this view.

The current evidence also supports the continued use of both subjective and objective measures as supplemental diagnostic procedures that are used as part of a comprehensive, integrated evaluation to more fully document vocal function and enhance the quality of diagnosis and treatment (Hillman et al., 1997). This view is still considered valid because the invasiveness of laryngeal imaging inherently limits its use, and noninvasive methods are needed for more frequent assessment (particularly for children) and use (e.g., biofeedback) during treatment, especially by SLPs. In addition, current image-based clinical assessments of vocal fold vibration (e.g., stroboscopy) are highly subjective and are not capable of reliably, accurately, and objectively characterizing the detailed impact of disorders on voice production; that is, the impact on underlying physiological mechanisms (e.g., associated air pressure, muscle activity, and straining), sound production (e.g., perturbation and aperiodic energy), and how the voice is ultimately perceived (e.g., voice quality) both by listeners and by the individual with the voice disorder (e.g., via functional measures).

Conclusion

Overall, the results of this systematic review used the CADE–M to provide evidence that selected acoustic, laryngeal imaging–based, auditory–perceptual, functional, and aerodynamic measures have the potential to be used as effective components in a clinical voice evaluation. The review, however, did not produce sufficient evidence on which to recommend a comprehensive set of methods for a standard clinical voice evaluation. There is clearly a pressing

need for high-quality research that is specifically designed to expand the evidence base for clinical voice assessment.

Acknowledgments

This systematic review was developed by the Ad Hoc Working Group for Clinical Voice Assessment under the auspices of Special Interest Group 3, Voice and Voice Disorders, of the American Speech-Language-Hearing Association (ASHA). The authors thank Leigh Deussing of ASHA for her support in developing the CADE–M form and for feedback during the review process. We thank Lauren Pecora and Kristin Slagle for their help with data transcription and transfer. We would also like to thank National Center for Evidence-Based Practice in Communication Disorders' staff members Tracy Schooling, Hillary Leech, and Rebecca Venediktov for their invaluable assistance with the literature review.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Arlington, VA: Author.

American Speech-Language-Hearing Association. (1998). *The roles of otolaryngologists and speech-language pathologists in the performance and interpretation of strobovideolaryngoscopy* [Relevant Paper]. Available from www.asha.org/policy.

American Speech-Language-Hearing Association. (2004). *Preferred practice patterns for the profession of speech-language pathology* [Preferred Practice Pattern]. Available from www.asha.org/policy.

American Speech-Language-Hearing Association. (2005). *Evidence-based practice in communication disorders* [Position Statement]. Available from www.asha.org/policy.

American Speech-Language-Hearing Association. (2008). *Loss to follow-up in early hearing detection and intervention* [Technical Report]. Available from www.asha.org/policy.

Arens, C., Reussner, D., Neubacher, H., Woenckhaus, J., & Glanz, H. (2006). Spectrometric measurement in laryngeal cancer. *European Archives of Oto-Rhino-Laryngology, 263*(11), 1001–1007.

Behrman, A. (2005). Common practices of voice therapists in the evaluation of patients. *Journal of Voice, 19,* 454–469.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry, 49,* 7–18.

Cammarota, G., Galli, J., Agostino, S. l., De Corso, E., Rigante, M., Cianci, R., … Gasbarrini, G. (2006). Accuracy of laryngeal examination during upper gastrointestinal endoscopy for premalignancy screening: Prospective study in patients with and without reflux symptoms. *Endoscopy, 38*(4), 376–381.

Coelho, C., Ylvisaker, M., & Turkstra, L. S. (2005). Nonstandardized assessment approaches for individuals with traumatic brain injuries. *Seminars in Speech and Language, 26,* 223–241.

Dejonckere, P. H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., … Woisard, V. (2001). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *European Archives of Oto-Rhino-Laryngology, 258,* 77–82.

Dice, G., & Shearer, W. M. (1973). Clinician's accuracy in detecting vocal nodules. *Language, Speech, and Hearing Services in Schools, 4,* 142–144.

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37,* 391–400.

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527–544.

Eadie, T. L., & Doyle, P. C. (2005). Classification of dysphonic voice: Acoustic and auditory-perceptual measures. *Journal of Voice: Official Journal of the Voice Foundation, 19*(1), 1–14.

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America, 110,* 2560–2566.

Godino-Llorente, J. I., & Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering, 51*(2), 380–384.

Hadjitodorov, S., Boyanov, B., & Teston, B. (2000). Laryngeal pathology detection by means of class-specific neural maps. *IEEE Transactions on Information Technology in Biomedicine, 4*(1), 68–73.

Hillman, R. E., Montgomery, W. W., & Zeitels, S. M. (1997). Appropriate use of objective measures of vocal function in the multidisciplinary management of voice disorders. *Current Opinion in Otolaryngology & Head and Neck Surgery, 5,* 172–175.

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., & Newman, C. W. (1997). The voice handicap index (VHI): Development and validation. *American Journal of Speech-Language Pathology, 6,* 66–70.

Jiang, J., & Stern, J. (2004). Receiver operating characteristic analysis of aerodynamic parameters obtained by airflow interruption: A preliminary report. *The Annals Of Otology, Rhinology, And Laryngology, 113*(12), 961–966.

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.

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine, 6,* 23.

Ma, E., Robertson, J., Radford, C., Vagne, S., El-Halabi, R., & Yiu, E. (2007). Reliability of speaking and maximum voice range measures in screening for dysphonia. *Journal of Voice, 21*(4), 397–406.

Mulherin, S. A., & Miller, W. C. (2002). Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine, 137,* 598–602.

Parsa, V., & Jamieson, D. G. (2000). Identification of pathological voices using glottal noise measures. *Journal of Speech, Language, and Hearing Research, 43,* 469–485.

Parsa, V., & Jamieson, D. G. (2001). Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research, 44*(2), 327–339.

Pribuisiene, R., Uloza, V., & Kupcinskas, L. (2008). Diagnostic sensitivity and specificity of laryngoscopic signs of reflux laryngitis. *Medicina* (Kaunas), *44*(4), 280–287.

Rees, C. J., Blalock, P. D., Kemp, S. E., Halum, S. L., & Koufman, J. A. (2007). Differentiation of adductor-type spasmodic dysphonia from muscle tension dysphonia by spectral analysis. *Otolaryngology – Head and Neck Surgery, 137*(4), 576–581.

Reid, M., Lachs, M. S., & Feinstein, A. R. (1995). Use of methodological standards in diagnostic test research: Getting better but still not good. *Journal of the American Medical Association, 274,* 645–651.

Rovirosa, A., Ascaso, C., Abellana, R., Martinez-Celdran, E., Ortega, A., Velasco, M., … Biete, A. (2008). Acoustic voice analysis in different phonetic contexts after larynx radiotherapy for T1 vocal cord carcinoma. *Clinical and Translational Oncology, 10*(3), 168–174.

Sackett, D. L., Rosenberg, W., Gray, J., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312,* 71–72.

Sackett, D., Straus, S., Richardson, W., Rosenberg, W., & Haynes, R. (2000). *Evidence-based medicine: How to practice and teach EBM*. New York, NY: Churchill Livingstone.

Schwartz, S. R., Cohen, S. M., Dailey, S. H., Rosenfeld, R. M., Deutsch, E. S., Gillespie, M. B., … Patel, M. M. (2009). Clinical practice guideline: Hoarseness (dysphonia). *Otolaryngology–Head and Neck Surgery, 141,* S1–S31.

Stojadinovic, A., Henry, L. R., Howard, R. S., Gurevich-Uvena, J., Makashay, M. J., Coppitt, G. L., … Solomon, N. P. (2008). Prospective trial of voice outcomes after thyroidectomy: Evaluation of patient-reported and clinician-determined voice assessments in identifying postthyroidectomy dysphonia. *Surgery, 143*(6), 732–742.

Titze, I. R. (1995). *Workshop on acoustic voice analysis: Summary statement*. Denver, CO: National Center for Voice and Speech.

Turkstra, L. S., Coelho, C., & Ylvisaker, M. (2005). The use of standardized tests for individuals with cognitive–communication disorders. *Seminars in Speech and Language, 26,* 215–222.

Umapathy, K., Krishnan, S., Parsa, V., & Jamieson, D. G. (2005). Discrimination of pathological voices using a time-frequency approach. *IEEE Transactions on Biomedical Engineering, 52*(3), 421–430.

Verdolini, K., Rosen, C., & Branski, R. C. (Eds.). (2006). *Classification manual for voice disorders–I, Special Interest Division 3, Voice and Voice Disorders, American Speech-Language Hearing Division*. Mahwah, NJ: Erlbaum.

Verikas, A., Gelzinis, A., Bacauskiene, M., & Uloza, V. (2006). Integrating global and local analysis of color, texture and geometrical information for categorizing laryngeal images. *International Journal of Pattern Recognition and Artificial Intelligence, 20*(8), 1187–1205.

Yorkston, K. M., Spencer, K., Duffy, J., Beukelman, D., Golper, L. A., Miller, R., … Sullivan, M. (2001). Evidence-based medicine and practice guidelines: Application to the field of speech-language pathology. *Journal of Medical Speech-Language Pathology, 9,* 243–256.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., & Glaze, L. E. (2011). Establishing validity of the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology, 20,* 14–22.

Appendix A

Electronic Databases Searched (Search Platform)

CINAHL (EBSCO)
Cochrane Library
ComDisDome (CSA)
Communication and Mass Media (EBSCO)
CRD Database
EBM Guidelines
Education Abstracts (EBSCO)
Education Research Complete (EBSCO)
ERIC
Health Source: Nursing/Academic Edition (EBSCO)
HighWire Press
Linguistics Language Behaviour Abstracts (CSA)
National Electronic Library for Health
National Rehabilitation Information Center (REHABDATA)
Neuroscience Abstracts (CSA)
OT Seeker
PEDro Physiotherapy Evidence Database
PsycARTICLES (EBSCO)
PsycBITE
Psychology and Behavioral Sciences Collection (EBSCO)
PsycINFO (EBSCO)
PubMed
Science Citation Index Expanded (ISI)
ScienceDirect
Social Sciences Citation Index (ISI)
Social Services Abstracts (CSA)
SUMSearch
Teacher Reference Center (EBSCO)
TRIP Database

Appendix B

Key Words

Voice measures

Case History: psychosocial, psychological attributes, personality traits, professional voice, non-professional voice, drugs, surgeries, voice use, reflux, symptoms, familial

Auditory–Perceptual Judgments: grade, hard glottal attack, breathiness, strain, hoarseness, rough, loudness, pitch, severity, CAPE-V, GRBAS, falsetto, fry, harsh, tremor, spasmodic, voice breaks, pitch breaks, aphonia, dysphonia, diplophonia, monotone, monoloudness, voice quality, voice profile analysis

Aerodynamic Measures: airflow (combined with voice, larynx, glottal), open quotient, DC flow, minimum flow, AC flow, peak flow, speed quotient, MFDR (maximum flow declination rate), average flow, airway resistance, glottal resistance, laryngeal airway resistance, subglottal pressure, PTP (phonation threshold pressure), circumferential mask, Aerophone, phonation quotient, S/Z ratio, maximum phonation time

Functional Measures: VHI (voice handicap index), VRQOL (voice related quality of life), pediatric VHI, VOISS, VHI-10, VAP, dysphonia severity index (DSI), FIMS for voice, VOS (voice outcomes survey), pediatric VRQOL, singing related quality of life

Acoustic Measures: jitter, shimmer, perturbation, perturbation quotient, relative amplitude perturbation, period perturbation quotient, amplitude perturbation quotient, voice (phonatory) breaks, subharmonic(s), tremor index, voice onset time, harmonic, H1/H2, formant, spectrum, short term spectrum, long term spectrum, fundamental frequency, range, intensity, decibel (dB), Hertz, voice range profile, Visi-Pitch, computerized speech lab (CSL), load, dosimeter, ambulatory phonation monitor, vocal load, timing, Type I, II, III, IV, aperiodic(ity), periodic, maximum phonation time, MPT, vocal efficiency, voice turbulence index, soft phonation index, phonetogram, frequency/intensity range profile, dysphonia severity index, cepstrum, cepstral peak prominence, TF32, CSpeech, Praat, Multidimensional voice program (MDVP), spectrogram, narrow band, vibration dose, distance dose, cycle dose, HNR (harmonic to noise), NHR (noise to harmonic), NNE (normalized noise energy), SNR (signal to noise ratio), wow, flutter, vocal attack time, F2 slope, spectral slope

Imaging: vocal fold, vocal cord, vocal fold imaging, laryngeal imaging, high-speed imaging, transnasal (larynx), high-speed videoendoscopy, fiberscopic (larynx), videostroboscopy, videoendoscopy (larynx), stroboscopy, videolaryngostroboscopy, laryngoscopy, rigid (larynx), oral rigid (larynx), transnasal flexible, nasoendoscopy, nasopharyngolaryngoscopy, kymography, videokymography, transillumination (larynx), ultrasound, laryngeal mirror, indirect laryngoscopy, direct laryngoscopy, vocal fold abduction, vocal fold adduction

Physical Exams: oral mechanism, extrinsic laryngeal palpation

Electroglottography (EGG): fundamental frequency (Fo, f0), open quotient, speed quotient, electroglottography, closed quotient

Voice disorder

voice, vocal, dysphonia

Diagnostic accuracy

correctly classify, classification accuracy, correct classification, classification error, classification rate, correctly classified, true positive classification, voice classification system, diagnostic accuracy, accurate diagnoses, accuracy diagnosis, accuracy predicting, prognosis accurate, positive predictive value, negative predictive value, correct assignment, correctly predicted, diagnoses modified, predicted correctly, receiver operating characteristics, ''ROC'', area under receiver operating characteristic curve, ''AROC'', sensitivity, sensitive, specificity, odds ratio, likelihood ratio, percentage correctly identified, percentage correctly classified, percentage correctly diagnosed, correctly classified, false positive, false negative, change diagnosis, alter diagnosis, percentage agreement

Additional search terms

Voice* OR vocal* OR dysphon*) AND (classify OR classification OR classified OR accuracy OR accurate OR predictive OR correct assignment OR correctly predicted OR diagnoses modified OR predicted correctly OR receiver operating characteristic* OR ''ROC'' OR under receiver operating characteristic curve OR ''AROC'' OR Sensitivity OR sensitive OR specificity OR odds ratio* OR likelihood ratio* OR percent* correctly identif* OR percent* correctly classif* OR percent* correctly diagnos* OR correctly classif* OR false positive OR false negative OR change diagnosis OR alter diagnosis OR percent* agreement

Appendix C (p. 1 of 3)

Modified Critical Appraisal of Diagnostic Evidence (CADE–M; Based on Dollaghan, 2007)

(1) Evaluator:

(2) Citation:

(3) Clinical Question:

Summarize the clinical question and then categorize it in terms of the assessment procedure being used to determine: 1) presence/absence, 2) nature (etiology), and/or 3) extent (severity) of a voice disorder.

(4) Describe Participant Characteristics:

Include a complete description of N, # of groups/diagnoses, # of females/males, Mean Age, SD, and range. If normals (i.e., normophonics) were used choose one of the following:

Normal status verified by endoscopy	Normal status **not** verified by endoscopy

(5) Describe Test Measure:

Describe new method used to classify/distinguish participants, and then categorize it as: case history, auditory–perceptual judgment, visual–perceptual judgment of endoscopic images, aerodynamic measure, functional measure, acoustic measure, image processing measure, physical exam, and/or electroglottographic measure.

(6) Describe Reference Standard/Measure (standard method used to classify participants):

If diagnoses of voice disorders were used as the reference, were these determined using standard medical procedures?

Yes	No ***OR*** not stated.

Appraisal Points

A. Study Design

Prospective Study	Retrospective Study

Indicate the type of design used in the study. The following are definitions of study designs. Prospective Study: A study designed to follow subjects forward in time, comparing a diagnostic test and reference standard. Retrospective Study: A study looking back in time, comparing the results of a diagnostic test and reference standard.

B. Reference Standard

Appropriate/reasonable reference standard used for comparison	Reference standard not appropriate/reasonable for comparison

The method to which a new diagnostic test or procedure is compared. The reference standard, often referred to as the ''gold standard,'' should be an existing test or diagnostic procedure whose accuracy is known and can reasonably be compared to the diagnostic test under investigation. In the absence of a ''gold standard'' with known accuracy, the reference standard is also considered the ''best available method for establishing the presence or absence of the target condition (p. 71)'' (Sackett et al., 1996). Indicate whether the reference standard was appropriate/reasonable, valid, and reliable for comparison to the diagnostic test under investigation.

Appendix C (p. 2 of 3)

Modified Critical Appraisal of Diagnostic Evidence (CADE–M; Based on Dollaghan, 2007)

C. Selection/Recruitment

Random or consecutive selection	Convenience sample/Hand-picked sample ***OR*** Not stated

Selection/recruitment refers to how subjects were chosen for the study. The criteria should be clearly defined to ensure that selection of subjects was not influenced by a likelihood of obtaining a particular result. Subjects should be selected either as a consecutive series or randomly from a clearly defined population.

D. Blinding

Assessors blinded when interpreting results of test and reference	Assessors not blinded when interpreting results of test and reference ***OR*** Not stated

Blinding refers to the practice of keeping investigators ignorant or blind to the results of each test being carried out. Investigators interpreting the results of one test should be blind to the results of the other test. Indicate whether the assessors were blinded when interpreting the test results or not blinded when interpreting the test results ***OR*** if blinding was not stated in the study.

E. Subjects

Subjects adequately described and similar to population in which test would be used with full spectrum of severity	Subjects not adequately described ***OR*** Subjects not similar to population in which test would be used with full spectrum of severity

The subjects in a study should include an appropriate spectrum of patients to whom the diagnostic test is applied in clinical practice. The subjects should be adequately described (e.g., age, sex, disease, setting) and the full spectrum of the disease should be considered (mild to severe). Indicate whether the subjects were adequately described and similar to population in which the test would be used with full spectrum of severity or if subjects were not adequately described ***OR*** if they were not similar to the population in which the test would be used. In a similar vein, when a diagnostic test is used to distinguish between two voice disordered groups (i.e., differential diagnosis), evidence should be provided that the two groups of interest display similar levels of severity such that the performance of the diagnostic test does not merely reflect uneven levels of disorder severity.

F. Avoidance of Verification Bias

Decision to perform reference standard independent of test results	Decision to perform reference standard not independent of test results ***OR*** Not stated

Ideally the diagnostic test under investigation and the reference standard should be performed on all subjects in the study to allow for the results to be compared. However, in some cases, the reference standard may not be carried out on all subjects, or an alternative to the reference standard is completed based on expense, invasiveness, etc. The decision to perform the reference standard should not be determined based on the results of the test under investigation. If the reference standard was not performed on all subjects, indicate whether this decision was independent of the results of the test under investigation or not independent of the test results ***OR*** if reason was not stated.

Appendix C (p. 3 of 3)

Modified Critical Appraisal of Diagnostic Evidence (CADE–M; Based on Dollaghan, 2007)

G. Results

Likelihood ratios reported or calculable	Likelihood ratios not reported or calculable
Was diagnostic accuracy adequate? LR+ > 10; LR– < .10	Was diagnostic accuracy inadequate? LR+ < 10; LR– > .10

The clinical usefulness of a diagnostic test is largely determined by the accuracy with which it identifies its target disorder. Likelihood ratios refer to how likely it is to obtain a given test result for a subject with the target disorder compared to the likelihood that that same result would be obtained in a subject without the target disorder. Sensitivity and specificity or raw data (i.e., 2×2 tables) are needed to obtain likelihood ratios. Therefore, at the very least, estimates of sensitivity and specificity, or raw data which permit calculation of sensitivity and specificity need to be provided. These data should stipulate the specific cutoff points (test scores/values) that were used to determine a positive test. In addition, some assessment of the performance/precision/accuracy of the diagnostic test is recommended following the general guidelines provided. This assessment is made understanding the inherent differences related to tests designed for diagnosis (i.e., determining if a disorder is present) versus differential diagnosis (i.e., distinguishing among competing diagnoses). In the space below, provide a brief ''summary'' of the diagnostic/test performance based on any estimates of diagnostic accuracy reported or calculated, i.e., sensitivity, specificity, likelihood ratios, PPV, NPV, AROC, or others. Use the cutoff criterion which provided the optimal test performance.

H. Follow up

Results reported on all subjects entered into study	Reasonable loss to follow up, < 20% of results not reported	Greater than 20% of results not reported

Results for all subjects entered into the study should be reported, including any for whom test results are unavailable for any reason. Studies in which a significant number of subjects do not have reported results should be noted. Indicate whether results were reported for all subjects entered into the study, < 20% of the results were not reported, or if greater than 20% of the results were not reported.

I. Validity of Findings

Compelling	Suggestive	Equivocal

J. Importance of Findings

Compelling	Suggestive	Equivocal

Sections I and J require an ''integrated judgment over the various appraisal criteria regarding whether the validity and importance of the external evidence were compelling (unarguable), suggestive (debatable on a few points, but leading to a consistent conclusion), or equivocal (debatable on so many points that unbiased and competent raters could reach opposite conclusions). If validity and importance are both compelling, there is serious reason to consider adding or adopting the test measure for use in diagnostic decision-making. If either validity or importance of external evidence is equivocal, no change to current diagnostic practice needs to be considered (although a change might be considered based upon the particular patient's characteristics or preferences). And if validity and importance of external evidence are both at least suggestive, different clinicians might well reach different decisions about whether to consider using the test measure for diagnosing a condition, p. 101 (Dollaghan, 2007).

K. Clinical Bottom Line

State simply whether the appraisal of the evidence does or does not support considering a change in one's current diagnostic approach—whether to adopt a new diagnostic measure or to stop using an inadequate diagnostic measure, p. 101 (Dollaghan, 2007).