Algorithm Verification and Concurrent Validity of a Web-Based Platform for Multiparametric Acoustic Voice Quality Indices☆

Jorge C. Lucero *Brasília, Brazil*

SUMMARY: Objectives. To evaluate whether established multiparametric acoustic voice quality indices, namely, the acoustic voice quality index (AVQI), acoustic breathiness index (ABI), cepstral spectral index of dysphonia (CSID), and smoothed cepstral peak prominence (CPPS), retain their concurrent validity with expert perceptual ratings when deployed through a free, web-based analysis platform.

Study design. Cross-sectional validation study using a publicly available voice database.

Methods. Voice samples from 290 speakers in the Perceptual Voice Qualities Database were analyzed using PhonaLab, a free web-based platform that implements these indices via Parselmouth (a Python interface to *Praat*). Sustained vowels and connected speech (consensus auditory-perceptual evaluation of voice [CAPE-V] sentences) were processed to extract AVQI, ABI, CSID, CPPS, and related parameters. Algorithm agreement with desktop *Praat* was assessed using Pearson correlations and intraclass correlation coefficients (ICCs). Spearman correlations with CAPE-V and GRBAS perceptual ratings were calculated. Receiver operating characteristic curve analysis was performed as a secondary, exploratory analysis.

Results. Algorithm agreement between PhonaLab and desktop *Praat* was strong to excellent ($r \geq 0.96$, ICC ≥ 0.94 for all primary indices). Correlations between acoustic indices and perceptual ratings were consistent with published validation studies: AVQI with CAPE-V Overall Severity ($r_s = 0.73$) and GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) Grade ($r_s = 0.75$); ABI with CAPE-V Breathiness ($r_s = 0.75$); CSID with GRBAS Grade ($r_s = 0.68$); CPPS from sustained vowels with CAPE-V Breathiness ($r_s = -0.69$). DeLong tests showed that ABI significantly outperformed single parameters for breathiness detection ($P < 0.001$).

Conclusions. Validated multiparametric acoustic indices retain their established concurrent validity with perceptual ratings when deployed through PhonaLab's web-based environment, supporting the feasibility of delivering research-validated acoustic analysis through browser-based platforms.

Keywords: Acoustic voice quality index–Acoustic breathiness index–Cepstral spectral index of dysphonia–Voice assessment–PVQD–Web-based analysis.

INTRODUCTION

Comprehensive voice evaluation requires a multidimensional approach integrating auditory-perceptual judgment, acoustic analysis, laryngoscopic examination, aerodynamic measures, and patient self-assessment.[1] Perceptual evaluation using the consensus auditory-perceptual evaluation of voice (CAPE-V)[2] and the Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) scale[3] remains the clinical gold standard but is inherently subjective, with documented variability that can limit reliability and complicate longitudinal monitoring.[4,5] Acoustic analysis offers an objective complement, providing quantifiable measures that can be standardized across time points and settings, as endorsed by the American Speech-Language-Hearing Association expert panel recommendation of cepstral peak prominence (CPP) as a robust correlate of dysphonia severity.[6]

Traditional acoustic parameters such as jitter, shimmer, and harmonics-to-noise ratio (HNR) present well-documented limitations: they require reliable fundamental frequency (F0) tracking, which becomes unreliable in moderately to severely dysphonic voices,[7,8] and show only weak to moderate correlations with perceptual ratings.[9,10] CPP, which does not require pitch tracking and applies to both sustained vowels and connected speech, has emerged as a more robust alternative.[11–13] In practice, the smoothed variant (smoothed cepstral peak prominence [CPPS]) implemented in *Praat*[14] is most commonly used in clinical research and is the version employed in this study; hereafter, CPPS refers specifically to this smoothed measure.

Recognition that even robust single parameters capture only one aspect of voice quality perception led to the development of composite indices. The acoustic voice quality index (AVQI)[15] combines six parameters from sustained vowel and connected speech, with strong concurrent validity demonstrated across numerous languages (pooled sensitivity 0.82, specificity 0.92 for version 03.01).[16] The acoustic breathiness index (ABI)[17] targets perceived breathiness specifically using nine weighted parameters, with correlations of $r_s = 0.75$–0.87 in validation studies.[18] The cepstral spectral index of dysphonia

Accepted for publication April 8, 2026.

☆ This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil. The funding source had no involvement in the study design, data collection, analysis, interpretation, writing, or decision to submit the article.

From the Department of Computer Science, University of Brasília, Brasília 70910-900, DF, Brazil. Email: lucero@unb.br

Journal of Voice, Vol xx, No xx, pp. xxx–xxx

0892-1997

https://doi.org/10.1016/j.jvoice.2026.04.009

(CSID)[12] combines CPP with spectral energy ratio measures, showing strong correlations with dysphonia severity (r = 0.81) and good diagnostic accuracy (area under the receiver operating characteristic curve [AUC] = 0.85) as a screening tool.[19]

Despite strong evidence supporting these indices, their uptake in routine clinical practice has been limited. Behrman[20] found that experienced voice therapists were significantly more likely to use subjective assessments than objective instrumental measures, including acoustic analysis. Nearly two decades later, the gap persists: Salgado et al[21] reported that 28% of surveyed clinicians lacked access to acoustic equipment, with postgraduate training and instrumentation access being significant predictors of whether acoustic measures were used. In a global multidisciplinary survey, Payten et al[22] found that only about half of responding speech-language pathologists routinely collected acoustic measures as part of initial voice evaluations, with notable inconsistencies relative to published protocols in equipment, voice samples collected, and types of measures obtained. A recent qualitative study[23] identified recurring themes among voice-specialized clinicians: collecting and analyzing acoustic data is time-consuming, the measures do not always directly inform therapy planning, yet they enable the most accurate longitudinal comparisons and support objective documentation of treatment outcomes. These converging findings suggest that practical barriers—including software cost, technical complexity, and time required for analysis—remain significant obstacles to acoustic assessment even as its clinical value is widely recognized. Commercial acoustic analysis packages typically cost several hundred to several thousand dollars, and while *Praat*[24] is freely available, calculating composite indices such as AVQI and ABI requires specialized scripts and technical expertise that many clinicians lack.[25,26]

Web-based platforms that implement validated algorithms with user-friendly interfaces represent a potential approach to reducing these barriers by eliminating software installation, script management, and the associated technical overhead. However, their ability to faithfully reproduce reference implementations must be empirically verified before they can be used with confidence. Implementation differences between software platforms—even those calling the same underlying signal processing code—can affect acoustic outputs through subtle variations in numerical precision, parameter settings, or processing pipelines, potentially altering clinical classifications.[13] Documenting the specific software versions and signal processing settings used in any acoustic analysis platform is therefore essential for reproducibility.[6] The primary objective of this study was to evaluate whether PhonaLab (`www.phonalab.com`), a free web-based platform developed by the author, faithfully deploys the AVQI, ABI, CSID, and CPPS algorithms and whether these indices retain their established concurrent validity with expert perceptual ratings when delivered through this platform. Specifically, we hypothesized that (1) PhonaLab's implementations would show strong agreement with desktop *Praat* (intraclass correlation coefficient [ICC] ≥ 0.90), and (2) correlations between acoustic indices and perceptual ratings from the PVQD would be consistent in magnitude and direction with those reported in published validation studies. As a secondary, exploratory objective, candidate clinical cutoff values for American English speakers were examined using receiver operating characteristic (ROC) curve analysis, with the ground truth defined by dichotomized perceptual ratings (CAPE-V ≥ 10 or GRBAS ≥ 0.5; see *statistical analysis* for rationale).

METHODS

Database

Voice recordings were obtained from the Perceptual Voice Qualities Database (PVQD), a publicly available collection of 296 speakers developed by Walden et al.[27] The PVQD contains high-quality audio recordings of sustained vowels and CAPE-V sentences from speakers with and without voice disorders across a broad range of dysphonia severity. Each recording was rated by at least three experienced voice clinicians (minimum 2 years experience) using both the CAPE-V and GRBAS scales, with established inter-rater and intra-rater reliability.[27] Perceptual ratings used in all analyses were the mean across all available raters for each speaker, as provided in the PVQD dataset. Because different raters evaluated different subsets of speakers in the PVQD design, inter-rater reliability was assessed using ICC(1, k) for consistency,[28] yielding values of 0.918 for CAPE-V Overall Severity, 0.827 for CAPE-V Breathiness, and 0.911 for GRBAS Grade.[27] The non-overlapping rater assignment precluded modeling individual rater effects analytically in secondary analyses; accordingly, the present study used the published mean ratings as the perceptual reference standard, consistent with other PVQD-based studies.[29,30]

Recordings were obtained using a head-mounted condenser microphone at 6 cm mouth-to-microphone distance, 16-bit quantization, and 44.1 kHz sampling rate using the Computerized Speech Lab (CSL; PENTAX Medical).

Sample selection

From the original 296 speakers, recordings were excluded due to missing sustained /a/ vowel segments or audible artifacts, yielding a final sample of 290 speakers for further analysis. The sample included speakers with voice disorders (n = 181) and vocally healthy controls (n = 109). From this set, the records of two speakers had missing CAPE-V ratings.

Audio segmentation

Each recording was manually segmented into two files: (1) a sustained /a/ vowel segment and (2) concatenated CAPE-V sentences representing connected speech. Segmentation was performed by the author using *Audacity* (version 3.7.5; Audacity Team, 2025). Given the clear silence intervals between speech tasks in the PVQD recordings,

segmentation boundaries were unambiguous in virtually all cases. Vowel boundaries were identified by visual and auditory inspection of the waveform and spectrogram. A limitation is that segmentation was performed by a single rater without formal reliability testing, though the clear task delineation minimized potential variability.

Acoustic analysis platform

All acoustic analyses were performed using PhonaLab. Its backend utilizes Parselmouth,[31] a Python library that embeds *Praat*'s C/C++ signal processing code within a Python interface, enabling server-side computation without requiring end users to install Praat or execute specialized scripts. Working within the Python ecosystem provides access to well-maintained libraries for statistical analysis, machine learning, and web deployment, facilitating reproducible research workflows.

The platform implements AVQI (version 03.01) and ABI according to the original specifications of Maryn et al[15] and Barsties v. Latoszek et al,[17] respectively. CSID for connected speech was calculated using the formula of Awan et al.[19] All algorithms use standard methods documented in the Praat manual and accessible through Parselmouth, enabling independent replication.

Three analysis protocols were applied: (1) sustained vowel analysis extracting jitter, shimmer, HNR, and CPPS; (2) multiparametric index calculation (AVQI, ABI) from combined vowel and speech samples; and (3) connected speech spectral analysis extracting CSID and Alpha Ratio.

Algorithm agreement with reference software

Prior to the main analysis, PhonaLab's implementations were compared against desktop *Praat* to verify computational fidelity. The same segmented audio files (sustained vowel and connected speech) used for the PhonaLab analysis were also analyzed using desktop *Praat* (version 6.4.43).[24] The official AVQI/ABI Praat script[32] was adapted for batch processing to generate AVQI, ABI, and individual acoustic parameter values for all speakers. For single-parameter vowel measures (jitter, shimmer, HNR, CPPS), the standard Praat voice report functions were used with identical settings. This parallel analysis of identical audio files enabled direct comparison of outputs between the two platforms.

Agreement was assessed using Pearson correlations, ICCs, mean bias, and 95% limits of agreement (LoA) as described in *statistical analysis*. This comparison serves as a technical verification that the web-based implementation faithfully reproduces the reference algorithms—a necessary precondition for the concurrent validity analysis that follows.

Because Parselmouth embeds Praat's C/C++ signal processing engine, strong agreement for individual acoustic parameters is expected by design. However, the web-based deployment introduces additional processing steps—including server-side file handling, audio format conversion, and Python-level orchestration of the analysis pipeline—that could introduce discrepancies. This comparison, therefore, serves not as an independent algorithmic validation but as an empirical verification that the end-to-end platform produces outputs consistent with the desktop reference.

Statistical analysis

All statistical analyses were performed in Python (version 3.11) using scipy (version 1.11), pingouin (version 0.5), and scikit-learn (version 1.3).

Primary analyses

Two primary hypotheses were tested. First, algorithm agreement between PhonaLab and desktop Praat was assessed using Pearson correlations, ICCs (two-way random-effects model, single measures, absolute agreement[28]), mean bias, and 95% LoA. Strong agreement was defined a priori as ICC ≥ 0.90.

Second, concurrent validity was evaluated by computing Spearman rank correlation coefficients (r_s) between acoustic measures and perceptual ratings, given the ordinal nature of GRBAS and the non-normal distribution of CAPE-V ratings. Bootstrap 95% confidence intervals (CI) were computed using 1000 resamples. Effect sizes were interpreted as: weak ($|r_s| < 0.30$), moderate ($0.30 \le |r_s| < 0.50$), strong ($0.50 \le |r_s| < 0.70$), and very strong ($|r_s| \ge 0.70$).[33] Bonferroni correction was applied for all primary correlation analyses. Statistical significance was set at $P = 0.05$. Concurrent validity was considered supported if the observed correlations were consistent in magnitude and direction with those reported in published validation studies of the same indices.

Exploratory analyses

As a secondary, exploratory objective, ROC curve analysis was performed to examine candidate clinical cutoff values for American English speakers. This analysis was motivated by the absence of published AVQI and ABI cutoffs derived from an American English sample using the PVQD. The ground truth for dichotomization was defined using perceptual rating thresholds: CAPE-V Overall Severity ≥10 versus < 10 for overall dysphonia, GRBAS Grade ≥0.5 versus < 0.5 for overall dysphonia, and CAPE-V Breathiness ≥10 versus < 10 for breathiness. The CAPE-V threshold of 10 was selected as a clinically meaningful boundary between normal variation and minimal perceptible deviance, consistent with clinical interpretation of the 100-point visual analog scale.[2] The GRBAS threshold of 0.5 represents the midpoint between normal (0) and minimal severity (1) on the averaged ordinal scale.

AUC was computed as the exact nonparametric statistic using scikit-learn, with bootstrap 95% CI calculated from 2000 resamples. Candidate cutoff values were identified using Youden's J index.[34] DeLong tests[35] were used to compare AUC values between correlated measures. To assess the stability of resubstitution estimates, 10-fold

stratified cross-validation was performed, repeated 100 times for stability. In each fold, the optimal cutoff was derived on the training set using Youden's J index, and sensitivity, specificity, and AUC were evaluated on the held-out test set. Cross-validated estimates are reported alongside the resubstitution values. These cutoffs remain exploratory and require external validation before clinical use.

Ethical considerations

The PVQD is a publicly available, de-identified database released under a Creative Commons license.[27] No institutional review board approval was required. All analyses were conducted in accordance with the Declaration of Helsinki.

RESULTS

Sample characteristics

The final sample characteristics are presented in Table 1. CAPE-V Overall Severity ratings ranged from 0.33 to 98.50 (mean = 28.85, standard deviation [SD] = 24.63), with GRBAS Grade from 0 to 3 (mean = 1.06, SD = 0.89). The distribution was skewed toward lower severity levels, consistent with original database characteristics.

Algorithm agreement

Agreement between PhonaLab and desktop Praat is presented in Table 2 and illustrated by Bland–Altman plots for AVQI and CSID in Figure 1. All primary indices showed strong to excellent agreement ($r \geq 0.96$, ICC ≥ 0.94). Small systematic biases were observed for AVQI (0.41 units) and ABI (0.37 units), as well as a larger offset for CSID (5.00 units). These biases may be attributable to effects in the Python implementation of Praat in Parselmouth, as well as to differences in Praat versions: Parselmouth (version 0.4.7) is based on *Praat* 6.1.38 (January 2021), whereas the desktop Praat version used was 6.4.43 (September 2025). Traditional parameters (jitter, shimmer, HNR) showed near-perfect agreement ($r > 0.99$).

TABLE 1.
Demographic and Clinical Characteristics of the Sample

Characteristic	Value
Sample	
Total PVQD speakers	296
Final sample	290
Available for AVQI/ABI	288
Age, years	
Mean (SD)	46.4 (21.8)
Range	14–93
Sex, n (%)	
Female	191 (65.9%)
Male	99 (34.1%)
Voice status, n (%)	
Voice disorder	181 (62.4%)
Normal/healthy	109 (37.6%)
Perceptual ratings, mean (SD) [n = 288]	
CAPE-V overall severity	28.85 (24.63)
CAPE-V breathiness	19.21 (20.04)
GRBAS grade	1.06 (0.89)

Abbreviations: ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CAPE-V, consensus auditory-perceptual evaluation of voice; GRBAS, grade, roughness, breathiness, asthenia, strain; PVQD, perceptual voice qualities database.

Descriptive statistics

Descriptive statistics for acoustic measures are presented in Table 3. Mean AVQI was 3.03 (SD = 2.28), ABI was 3.71 (SD = 2.00), and CSID was 1.97 (SD = 41.60). Mean CPPS from sustained vowels was 13.40 dB (SD = 4.14). In some cases of the single parameters, measures could not be obtained due to very short vowels or errors in the detection of voice periods.

Correlations between acoustic measures and perceptual ratings

Spearman correlations between acoustic measures and the three primary perceptual criteria (CAPE-V Overall Severity, CAPE-V Breathiness, and GRBAS Grade) are presented in Table 4. All correlations were statistically significant ($P < 0.001$) after Bonferroni correction, except for Alpha Ratio, which showed only weak associations ($|r_s| = 0.15$–0.28).

AVQI demonstrated very strong correlations with GRBAS Grade ($r_s = 0.75$; 95% CI: 0.69–0.79) and strong correlations with CAPE-V Overall Severity ($r_s = 0.73$; 95% CI: 0.67–0.78) and CAPE-V Breathiness ($r_s = 0.72$). ABI showed its strongest correlation with its target dimension, CAPE-V Breathiness ($r_s = 0.75$; 95% CI: 0.69–0.80), with strong but lower correlations with Overall Severity and GRBAS Grade (both $r_s = 0.66$). This pattern confirms ABI's specificity for breathiness relative to the more general AVQI. CSID demonstrated strong correlations with GRBAS Grade ($r_s = 0.68$) and Overall Severity ($r_s = 0.67$). Among single parameters, CPPS from sustained vowels showed its strongest correlation with CAPE-V Breathiness ($r_s = -0.69$), while jitter ($r_s = 0.54$–0.67), shimmer ($r_s = 0.58$–0.69), and HNR ($r_s = -0.55$ to -0.66) showed strong, broadly similar correlations across perceptual dimensions.

To examine whether acoustic–perceptual associations varied across the dysphonia severity continuum, Spearman correlations were computed after stratifying speakers into three groups based on CAPE-V Overall Severity: normal-to-minimal (< 20, $n = 149$), mild-to-moderate (20–50, $n = 80$), and moderate-to-severe (> 50, $n = 59$; Table 5). As expected, within-stratum correlations were attenuated relative to full-sample values due to range restriction. However, the pattern of associations was preserved: AVQI maintained its strongest correlations with GRBAS Grade, and ABI retained its specificity for CAPE-V Breathiness across all severity levels, with notably strong correlations even within the moderate-to-severe stratum ($r_s = 0.72$).

TABLE 2.
Agreement between Phonalab and Desktop Praat for Acoustic Measures (N = 290)

Measure	*r*	ICC (2,1)	Bias	95% LoA	MAD
Jitter local, %	0.995	0.995	0.01	[– 0.24, 0.27]	0.03
Shimmer local, %	0.998	0.997	0.00	[– 0.63, 0.64]	0.07
HNR, dB	0.999	0.999	– 0.01	[– 0.63, 0.61]	0.03
CPPS, dB	0.980	0.979	– 0.13	[– 1.82, 1.56]	0.44
Alpha Ratio, dB	0.981	0.981	– 0.05	[– 1.69, 1.60]	0.05
AVQI	0.960	0.945	0.41	[– 0.84, 1.66]	0.46
ABI	0.972	0.956	0.37	[– 0.55, 1.29]	0.42
CSID	0.992	0.984	5.00	[– 5.95, 15.96]	5.55

Abbreviations: ABI, acoustic breathiness index; AVQI, acoustic voice quality index; Bias, mean difference (PhonaLab – Praat); CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; HNR, harmonics-to-noise ratio; ICC, intraclass correlation coefficient (two-way random, single measures, absolute agreement); LoA, limits of agreement; MAD, mean absolute difference; r, Pearson correlation.

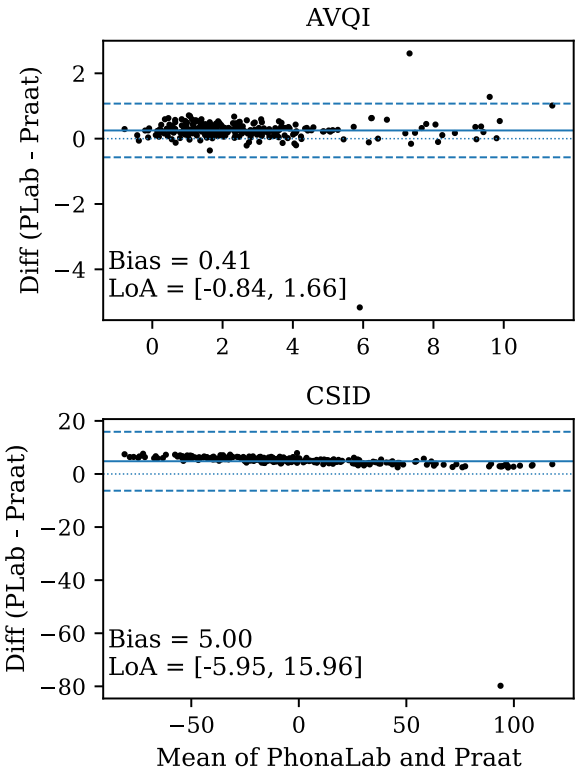

FIGURE 1. Bland–Altman plots comparing AVQI (left) and CSID (right) values computed by PhonaLab and desktop *Praat*. Solid lines indicate mean bias; dashed lines indicate 95% limits of agreement. AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; LoA, limits of agreement.

Exploratory diagnostic accuracy

The following ROC analyses were performed as a secondary, exploratory objective and were not part of the primary hypothesis testing. Results should be interpreted as hypothesis-generating.

ROC analysis evaluated the discriminative ability of acoustic measures for identifying dysphonic voices (CAPE-V Overall Severity ≥10) and breathy voices (CAPE-V Breathiness ≥10). Table 6 presents AUC values and candidate cutoffs for the primary indices; ROC curves for the four main indices are shown in Figure 2.

TABLE 3.
Descriptive Statistics for Acoustic Measures (N = 290).

Measure	Mean	SD	Min	Max
Multiparametric indices				
AVQI	3.03	2.28	– 0.64	10.95
ABI	3.71	2.00	– 0.01	9.45
CSID	1.97	41.60	– 77.90	119.70
Single-parameter measures (sustained vowel)				
F0, Hz	174.50	62.20	75.82	413.74
Jitter local, %	1.09	1.33	0.18	9.37
Shimmer local, %	5.10	4.56	0.93	22.53
HNR, dB	20.48	7.12	– 1.29	32.91
CPPS, dB	13.40	4.14	2.39	22.33

Abbreviations: ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; HNR, harmonics-to-noise ratio.

TABLE 4.
Spearman Correlations (r_s) between Acoustic Measures and Perceptual Ratings (N = 288).

Acoustic Measure	CAPE-V Severity	CAPE-V Breathiness	GRBAS Grade
AVQI	0.73	0.72	0.75
ABI	0.66	0.75	0.66
CSID	0.67	0.61	0.68
CPPS	– 0.63	– 0.69	– 0.62
HNR	– 0.66	– 0.55	– 0.66
Jitter local	0.67	0.54	0.67
Shimmer local	0.68	0.58	0.69

All $P < 0.001$ after Bonferroni correction.
Abbreviations: ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; CAPE-V, consensus auditory-perceptual evaluation of voice; GRBAS, grade, roughness, breathiness, asthenia, strain; HNR, harmonics-to-noise ratio.

AVQI yielded AUC = 0.825 for CAPE-V Severity, with a candidate cutoff of > 1.96 (sensitivity 73%, specificity 78%). ABI achieved the highest AUC among all indices for its target dimension (AUC = 0.862 for CAPE-V Breathiness), with a

TABLE 5.
Spearman Correlations (r_s) between Acoustic Indices and Perceptual Ratings, Stratified by CAPE-V Overall Severity

	Full Sample	Normal/Minimal	Mild–Moderate	Moderate–Severe
CAPE-V severity	n = 288	n = 149	n = 80	n = 59
AVQI	0.73‡	0.28‡	0.49‡	0.51‡
ABI	0.66‡	0.22†	0.44‡	0.50‡
CSID	0.67‡	0.24†	0.40‡	0.47‡
CPPS	−0.63‡	−0.27‡	−0.36‡	−0.45‡
CAPE-V breathiness	n = 288	n = 149	n = 80	n = 59
AVQI	0.72‡	0.42‡	0.34†	0.64‡
ABI	0.75‡	0.50‡	0.55‡	0.72‡
CSID	0.61‡	0.22†	0.29†	0.53‡
CPPS	−0.69‡	−0.49‡	−0.43‡	−0.63‡
GRBAS grade	n = 288	n = 149	n = 80	n = 59
AVQI	0.75‡	0.36‡	0.46‡	0.46‡
ABI	0.66‡	0.27‡	0.37‡	0.44‡
CSID	0.68‡	0.29‡	0.44‡	0.46‡
CPPS	−0.62‡	−0.31‡	−0.24*	−0.41†

Normal/Minimal: CAPE-V Overall Severity < 20; Mild–Moderate: 20–50; Moderate–Severe: > 50. Within-stratum correlations are attenuated relative to the full sample due to range restriction.

Abbreviations: ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; CAPE-V, consensus auditory-perceptual evaluation of voice; GRBAS, grade, roughness, breathiness, asthenia, strain.

* $P < 0.05$.

† $P < 0.01$.

‡ $P < 0.001$.

TABLE 6.
Exploratory ROC Analysis: Resubstitution and Cross-validated Diagnostic Accuracy

		Resubstitution				10-Fold Cross-Validation		
Measure	Reference Standard	AUC	Cutoff	Sens	Spec	AUC	Sens	Spec
AVQI	CAPE-V severity ≥ 10	0.825	> 1.96	73.4%	78.4%	0.825	73.3%	75.4%
AVQI	GRBAS grade ≥ 0.5	0.838	> 1.96	78.3%	75.3%	0.839	72.9%	74.8%
ABI	CAPE-V breathiness ≥ 10	0.862	> 2.97	84.3%	74.4%	0.861	80.0%	75.0%
ABI	GRBAS breathiness ≥ 0.5	0.850	> 3.75	80.3%	74.7%	0.852	73.9%	74.4%
CSID	CAPE-V severity ≥ 10	0.790	> − 3.60	61.7%	81.1%	0.790	62.7%	72.3%
CPPS	CAPE-V severity ≥ 10	0.796	< 13.33	58.4%	87.8%	0.795	57.9%	83.6%

Resubstitution: cutoff derived and evaluated on the full sample using Youden's J index. Cross-validation: 10-fold stratified, repeated 100 times; cutoff derived on training folds, evaluated on held-out test fold. Cross-validation values are means across all folds and repeats. $n = 288$ for all analyses. Cutoff values are from resubstitution and should be considered hypothesis-generating; external validation is required before clinical use.

Abbreviations: AUC, area under the ROC curve; ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; CAPE-V, consensus auditory-perceptual evaluation of voice; GRBAS, grade, roughness, breathiness, asthenia, strain; Sens, sensitivity; Spec, specificity.

candidate cutoff of > 2.97 (sensitivity 84%, specificity 74%). CSID showed AUC = 0.790, with moderate sensitivity (62%) and high specificity (81%). CPPS from sustained vowels yielded AUC = 0.796 with a candidate cutoff of < 13.33 dB (sensitivity 58%, specificity 88%). Among single parameters, shimmer achieved the highest AUC for dysphonia detection (0.836), followed by HNR (0.816) and jitter (0.812).

Cross-validated diagnostic accuracy estimates closely replicated the resubstitution values (Table 6). AUC optimism was negligible across all indices (≤0.002), indicating that the resubstitution estimates were not inflated by overfitting. Cross-validated sensitivity and specificity showed modest attenuation relative to resubstitution values (typically 2–5% points), consistent with cutoff derivation on smaller training sets. DeLong tests (Table 7) compared the discriminative performance of multiparametric indices against single parameters and against each other. For breathiness detection, ABI significantly outperformed HNR ($\Delta AUC = 0.097$, $P < 0.001$), supporting the added value of a breathiness-specific composite index. The difference between ABI and CPPS was not statistically significant ($\Delta AUC = 0.019$, $P = 0.209$).

For overall dysphonia detection, AVQI did not significantly outperform any single parameter, including CSID ($P = 0.150$), shimmer ($P = 0.677$), HNR ($P = 0.751$), or jitter ($P = 0.683$). When evaluated against GRBAS Grade, the AVQI–CSID difference reached significance ($\Delta AUC = 0.049$, $P = 0.031$). The sample may have been

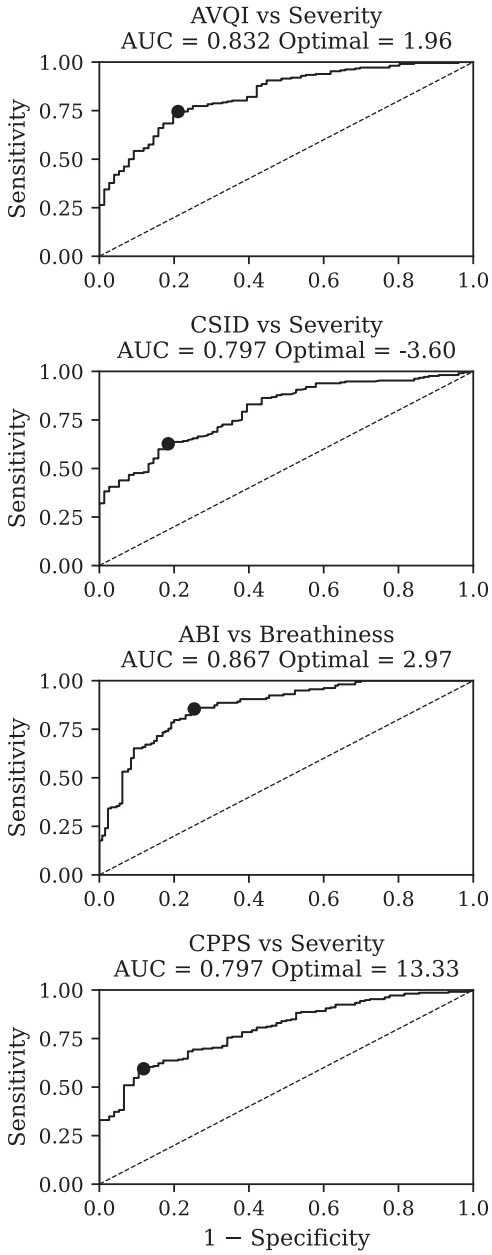

FIGURE 2. Receiver operating characteristic (ROC) curves for AVQI, CSID (speech), ABI, and CPPS in discriminating speakers with clinically relevant dysphonia. Clinically relevant deviation was defined as CAPE-V overall severity ≥10 for AVQI, CSID, and CPPS, and CAPE-V Breathiness ≥10 for ABI. The optimal cutoff for each measure was determined using Youden's J index. Area under the curve (AUC) values with bootstrap 95% confidence intervals are shown in each panel. ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; CAPE-V, consensus auditory-perceptual evaluation of voice.

underpowered to detect small AUC differences between highly correlated, effective tests—a common limitation when comparing two well-performing ROC curves. AVQI may nonetheless offer practical advantages through its single interpretable score integrating information from two speech tasks, even if its discriminative accuracy does not significantly exceed that of well-chosen single parameters in this sample.

DISCUSSION

Summary of findings

This study evaluated whether established multiparametric acoustic indices retain their concurrent validity with expert perceptual ratings when deployed through a web-based platform. Two complementary findings support this conclusion. First, PhonaLab's implementations showed strong to excellent agreement with desktop Praat ($r \geq 0.96$, ICC ≥ 0.94), confirming faithful reproduction of the reference algorithms. Second, the observed correlations between indices and perceptual ratings—AVQI with severity (r_s = 0.73–0.75), ABI with breathiness (r_s = 0.75), CSID with severity (r_s = 0.67–0.68), CPPS with breathiness (r_s = −0.69)—were consistent with published validation studies using commercial and desktop software. This study does not claim to re-validate the indices themselves, whose validity is well established; rather, it demonstrates that these validated algorithms can be successfully deployed in a new computational environment without compromising their concurrent validity.

Comparison with published literature

The AVQI–severity correlation (r_s = 0.73–0.75) falls within the range reported across 18 studies (r = 0.67–0.88) in the systematic review by Jayakumar and Benoy,[16] with our mid-range values possibly reflecting the heterogeneous PVQD sample. AUC values of 0.825–0.838 similarly align with published data.

The ABI–breathiness correlation (r_s = 0.75) is consistent with published values (r_s = 0.75–0.89) [17,18]. CSID correlations with perceived severity (r_s = 0.67–0.68) were moderate and lower than the r_s = 0.81 reported by Awan et al[12] for continuous speech, likely reflecting both the heterogeneous PVQD sample and the systematic offset between PhonaLab and Praat implementations (5.00 units), which means CSID cutoffs may not be directly transferable across platforms.

Severity-stratified analyses (Table 5) confirmed that the pattern of acoustic–perceptual associations was consistent across severity levels, with ABI retaining particularly strong breathiness-specific correlations within each stratum. The attenuation of within-stratum correlations relative to full-sample values reflects range restriction rather than diminished index performance.

The PVQD has been used in several recent studies, enabling direct comparison. Cantor-Cutiva et al[29] identified optimal parameter combinations for voice disorder screening, and Wischhoff et al[30] validated nonlinear chaos parameters in a PVQD subset (n = 55), reporting correlations with GRBAS Grade of $|r|$ = 0.71–0.72—comparable to the present findings. This consistency across studies supports the reliability of both the PVQD ratings and the acoustic methods.

TABLE 7.
DeLong Test Comparing Diagnostic Accuracy Values between Acoustic Measures

Comparison	AUC_1	AUC_2	Δ AUC	SE	*z*	*P*
Overall dysphonia (CAPE-V severity ≥ 10)						
AVQI vs HNR	0.825	0.816	+0.009	0.028	0.32	0.751
AVQI vs Jitter	0.825	0.812	+0.013	0.031	0.41	0.683
AVQI vs Shimmer	0.825	0.836	-0.011	0.026	-0.42	0.677
AVQI vs CSID	0.825	0.790	+0.034	0.024	1.44	0.150
Overall dysphonia (GRBAS grade ≥ 0.5)						
AVQI vs CSID	0.838	0.789	+0.049	0.023	2.16	0.031*
Breathiness (CAPE-V breathiness ≥ 10)						
ABI vs HNR	0.862	0.765	+0.097	0.028	3.49	< 0.001*
ABI vs CPPS	0.862	0.843	+0.019	0.015	1.26	0.209

AUC_1 = first measure; AUC_2 = second measure.
Abbreviations: AUC, area under the ROC curve; ABI, acoustic breathiness index; AVQI, acoustic voice quality index; CSID, cepstral spectral index of dysphonia; CPPS, smoothed cepstral peak prominence; CAPE-V, consensus auditory-perceptual evaluation of voice; GRBAS, grade, roughness, breathiness, asthenia, strain; SE, standard error.
* $P < 0.05$.

Comparative diagnostic accuracy

DeLong tests revealed that ABI significantly outperformed HNR for breathiness detection, validating the rationale for developing perceptually-targeted multiparametric indices. The difference between ABI and CPPS did not reach significance ($P = 0.209$), suggesting that CPPS alone captures much of the breathiness-related acoustic information.

For overall dysphonia, AVQI did not significantly outperform single parameters when CAPE-V Severity was used as the reference standard, though the AVQI–CSID difference reached significance when evaluated against GRBAS Grade ($P = 0.031$). The sample may have been underpowered to detect small AUC differences between highly correlated, effective tests—a common limitation when comparing two well-performing ROC curves. AVQI may nonetheless offer practical advantages through its single interpretable score integrating information from two speech tasks, even if its discriminative accuracy does not significantly exceed that of well-chosen single parameters in this sample.

Implementation reproducibility in acoustic voice analysis

Because PhonaLab accesses Praat's signal processing code through Parselmouth, strong agreement at the level of individual acoustic parameters might be expected by design. The agreement analysis in this study was therefore intended as a technical verification of the deployment pipeline, not as an independent algorithmic validation—the concurrent validity analysis with perceptual ratings serves that independent function. Nonetheless, the present findings illustrate that even implementations sharing the same core engine can produce clinically meaningful systematic offsets: most notably, the 5.00-unit CSID bias attributable to version-dependent numerical differences between Parselmouth (based on *Praat* 6.1.38) and desktop *Praat* (6.4.43). More broadly, acoustic voice analysis is increasingly used for clinical decision-making, yet the same nominal algorithm can produce different outputs depending on the software platform, library version, and processing pipeline.[13] These findings underscore the importance of documenting software versions in acoustic studies and of empirically verifying cross-platform equivalence rather than assuming it. Ensuring algorithmic consistency across software environments is essential for reproducible research and for reducing barriers to standardized clinical use of acoustic voice measures.

Limitations

Several limitations should be considered. First, the PVQD recordings were obtained using high-quality equipment in controlled conditions. The key question facing clinicians—whether comparable results would be obtained from smartphone recordings in acoustically variable settings—cannot be answered by this study. A logical next step is to evaluate these tools using recordings from typical clinical and telehealth environments.

Second, PVQD perceptual ratings used a 100-point visual analog scale without severity anchors, which may affect comparability with standard CAPE-V procedures. However, the strong correlations between CAPE-V and GRBAS ratings reported for this database[27] suggest adequate construct validity of the perceptual measures.

Third, the cross-sectional design precludes evaluation of sensitivity to change, essential for treatment outcome monitoring.

Fourth, although cross-validation confirmed negligible optimism in the candidate cutoff values, these remain exploratory and require external validation in an independent sample before clinical adoption.

Fifth, the CSID offset between platforms (5.00 units) means cutoffs established in one implementation may not transfer directly to another. More broadly, the version-dependent differences between Parselmouth (based on Praat 6.1.38) and current desktop *Praat* (6.4.43) underscore

the importance of documenting software versions in acoustic analysis studies.

Sixth, audio segmentation was performed by a single rater without formal reliability testing, though the clear silence intervals between tasks minimized potential variability.

Finally, Youden's J index maximizes the sum of sensitivity and specificity, which may not be optimal for all clinical contexts. In screening applications, lower cutoffs favoring sensitivity may be preferred. More broadly, all analyses relied on a single database (PVQD). While the PVQD provides a well-characterized, heterogeneous sample with established perceptual ratings, generalizability to other populations, recording conditions, and clinical settings remains to be demonstrated through external validation.

Future directions

Prospective validation of sensitivity to change following intervention would strengthen the clinical utility evidence. Evaluation with smartphone-recorded audio would test ecological validity for the intended use case. External validation of candidate cutoffs in an independent sample is needed before clinical adoption. Development of indices for additional perceptual dimensions (roughness, strain) would enable more comprehensive assessment.

CONCLUSIONS

This study demonstrates that established multiparametric acoustic indices (AVQI, ABI, CSID, CPPS) retain their concurrent validity with expert perceptual ratings when deployed through PhonaLab, a free web-based platform. Algorithm agreement with desktop Praat was strong to excellent, and the observed correlations with perceptual ratings were consistent with published validation studies. ABI provided significantly better diagnostic accuracy than single parameters for breathiness detection, while AVQI showed comparable accuracy to individual measures for overall dysphonia.

These findings support the feasibility of delivering research-validated acoustic analysis through browser-based platforms without compromising scientific validity. Candidate cutoff values were identified as an exploratory secondary analysis and require external validation. Further research is needed to evaluate sensitivity to change, performance with clinically-obtained recordings, and independent validation of cutoff thresholds.

Data and code availability

The PVQD is publicly available at https://data.mendeley.com/datasets/9dz247gnyb.[27]. PhonaLab is freely accessible at www.phonalab.com. The platform's acoustic analysis backend uses Parselmouth version 0.4.7,[31] which embeds the Praat 6.1.38 signal processing engine. Desktop Praat version 6.4.43 was used for reference comparisons. Statistical analyses were performed in Python 3.11 using scipy 1.11, pingouin 0.5, and scikit-learn 1.3. AVQI (version 03.01) and ABI were implemented according to the original specifications of Maryn et al,[15] and Barsties v. Latoszek et al,[17] respectively, using the standard Praat algorithms accessible through Parselmouth. CSID for connected speech was calculated using the formula of Awan et al.[19] Key analysis settings for CPPS extraction follow the official AVQI/ABI script:[32] PowerCepstrogram generation with pitch floor 60 Hz, time step 0.002 seconds, maximum frequency 5000 Hz, and pre-emphasis from 50 Hz; CPPS computed with time averaging window 0.01 s, quefrency averaging window 0.001 seconds, peak search quefrency range 60–330 Hz, parabolic interpolation, straight trend line with robust fit, and no tilt subtraction before smoothing. Statistical analysis scripts used in this study are available from the corresponding author upon reasonable request.

Conflict of interest statement

The author is the developer and sole operator of PhonaLab. The platform is provided free of charge with no commercial revenue at the time of this study. The analyses were not preregistered. To mitigate potential bias, this study used a publicly available database with published perceptual ratings, employed standard statistical methods, and compared results against an established reference implementation (desktop Praat).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the author used Claude (Anthropic) to assist with statistical code development, manuscript drafting, and editorial revision. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Acknowledgments

The author thanks the developers of the Perceptual Voice Qualities Database (PVQD) for making this resource publicly available, and Dr. Youri Maryn and Dr. Ben Barsties v. Latoszek for making the AVQI and ABI Praat scripts available to the research community.

References

1. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating

the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Oto Rhino Laryngol.* 2001;258:77–82. https://doi.org/10.1007/s004050000299.
2. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol.* 2009;18:124–132. https://doi.org/10.1044/1058-0360(2008/08-0017).
3. Hirano M. *Clinical Examination of Voice.* Vienna: Springer-Verlag; 1981.
4. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res.* 1993;36:21–40. https://doi.org/10.1044/jshr.3601.21.
5. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice.* 2006;20:527–544. https://doi.org/10.1016/j.jvoice.2005.08.007.
6. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *Am J Speech Lang Pathol.* 2018;27:887–905. https://doi.org/10.1044/2018_ajslp-17-0009.
7. Titze IR, Workshop on Acoustic Voice Analysis: Summary Statement, National Center for Voice and Speech, Denver, CO, 1995.
8. Jiang JJ, Zhang Y, McGilligan C. Chaos in voice, from modeling to measurement. *J Voice.* 2006;20:2–17. https://doi.org/10.1016/j.jvoice.2005.01.001.
9. Heman-Ackah YD, Michael DD, Baroody MM, et al. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Oto Rhinol Laryngol.* 2003;112:324–333. https://doi.org/10.1177/000348940311200406.
10. Maryn Y, Dick C, Vandenbruaene C, Vauterin T, Jacobs T. Spectral, cepstral, and multivariate exploration of tracheoesophageal voice quality in continuous speech and sustained vowels. *Laryngoscope.* 2009;119:2384–2394. https://doi.org/10.1002/lary.20620.
11. Heman-Ackah YD, Michael DD, Goding Jr GS. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice.* 2002;16:20–27. https://doi.org/10.1016/s0892-1997(02)00067-x.
12. Awan SN, Roy N, Jetté ME, Meltzner GS, Hillman RE. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clin Linguist Phonet.* 2010;24:742–758. https://doi.org/10.3109/02699206.2010.492446.
13. Murton O, Hillman R, Mehta D. Cepstral peak prominence values for clinical voice evaluation. *Am J Speech Lang Pathol.* 2020;29:1596–1607. https://doi.org/10.1044/2020_ajslp-20-00001.
14. Maryn Y, Weenink D. Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. *J Voice.* 2015;29:35–43. https://doi.org/10.1016/j.jvoice.2014.06.015.
15. Maryn Y, De Bodt M, Roy N. The acoustic voice quality index: toward improved treatment outcomes assessment in voice disorders. *J Commun Disord.* 2010;43:161–174. https://doi.org/10.1016/j.jcomdis.2009.12.004.
16. Jayakumar T, Benoy JJ. Acoustic voice quality index (avqi) in the measurement of voice quality: a systematic review and meta-analysis. *J Voice.* 2024;38:1055–1069. https://doi.org/10.1016/j.jvoice.2022.03.018.
17. Barsties v. Latoszek B, Maryn Y, Gerrits E, De Bodt M. The acoustic breathiness index (ABI): a multivariate acoustic model for breathiness. *J Voice.* 2017;31:511.e11–511.e27. https://doi.org/10.1016/j.jvoice.2016.11.017.
18. Barsties v. Latoszek B, Kim G-H, Delgado Hernández J, et al. The validity of the acoustic breathiness index in the evaluation of breathy voice quality: a meta-analysis. *Clin Otolaryngol.* 2020;46:31–40. https://doi.org/10.1111/coa.13629.
19. Awan SN, Roy N, Zhang D, Cohen SM. Validation of the cepstral spectral index of dysphonia (CSID) as a screening tool for voice disorders: development of clinical cutoff scores. *J Voice.* 2016;30:130–144. https://doi.org/10.1016/j.jvoice.2015.04.009.
20. Behrman A. Common practices of voice therapists in the evaluation of patients. *J Voice.* 2005;19:454–469. https://doi.org/10.1016/j.jvoice.2004.08.004.
21. Salgado S, Schils SA, Childes JM, Crino C, Palmer AD. Current practices in the assessment of voice: a comparison of providers across different clinical settings. *J Voice.* 2024. https://doi.org/10.1016/j.jvoice.2024.08.007. In press.
22. Payten CL, Weir KA, Madill CJ. Investigating current clinical practice in assessment and diagnosis of voice disorders: a cross-sectional multidisciplinary global web survey. *Int J Lang Commun Disord.* 2024;59:2786–2805. https://doi.org/10.1111/1460-6984.13116.
23. Madoule MD, Marks KL, Nagle KF, et al. Qualitative analysis of speech-language pathologists' voice evaluation practices and perspectives. *Am J Speech Lang Pathol.* 2025;34:2244–2259. https://doi.org/10.1044/2025_ajslp-24-00417.
24. Boersma P, Weenink D, Praat: Doing phonetics by computer [computer program].Version 6.4.43; 2025. Retrieved 23 October 2025 from ⟨http://www.praat.org/⟩.
25. Lee JM, Roy N, Peterson E, Merrill RM. Comparison of two multiparameter acoustic indices of dysphonia severity: the acoustic voice quality index and cepstral spectral index of dysphonia. *J Voice.* 2018;32:515.e1–515.e13. https://doi.org/10.1016/j.jvoice.2017.06.012.
26. Maryn Y. Practical acoustics in clinical voice assessment: a praat primer. *Perspect ASHA Special Interest Groups.* 2017;2:14–32. https://doi.org/10.1044/persp2.sig3.14.
27. Walden PR. Perceptual voice qualities database (PVQD): database characteristics. *J Voice.* 2022;36:875.e15–875.e23. https://doi.org/10.1016/j.jvoice.2020.10.001.
28. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull.* 1979;86:420–428. https://doi.org/10.1037/0033-2909.86.2.420.
29. Cantor-Cutiva LC, Ramani SA, Walden PR, Hunter EJ. Screening of voice pathologies: identifying the predictive value of voice acoustic parameters for common voice pathologies. *J Voice.* 2026;40:812–819. https://doi.org/10.1016/j.jvoice.2023.12.005.
30. Wischhoff OP, Gouraram V, Chumbley TJ, Liu B, Jiang JJ. Auditory-perceptual validation of acoustic chaos parameters in healthy and dysphonic voices. *J Speech Lang Hear Res.* 2025;68:5212–5225. https://doi.org/10.1044/2025_jslhr-25-00155.
31. Jadoul Y, Thompson B, de Boer B. Introducing parselmouth: a python interface to Praat. *J Phonet.* 2018;71:1–15. https://doi.org/10.1016/j.wocn.2018.07.001.
32. Barsties B, Maryn Y. External validation of the acoustic voice quality index version 03.01 with extended representativity. *Ann Otol Rhinol Laryngol.* 2016;125:571–583. https://doi.org/10.1177/0003489416636131.
33. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
34. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3:32–35.10.1002/1097-0142(1950)3:1 < 32::aid-cncr2820030106 > 3.0.co;2-3.
35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837–845. https://doi.org/10.2307/2531595.