

Research Article

Pediatric Auditory-Perceptual Evaluation of Voice Part I: Voice Stimuli for Children

Robert Brinton Fujiki,^a  Rebecca Johnson,^b Sijia Huang,^c and Susan L. Thibeault^b 

^aDepartment of Otolaryngology—Head and Neck Surgery, Indiana University School of Medicine, Indianapolis ^bDepartment of Otolaryngology—Head and Neck Surgery, University of Wisconsin–Madison ^cDepartment of Applied Psychology in Education and Research Methodology, Indiana University Bloomington

ARTICLE INFO

Article History:

Received November 12, 2025

Revision received March 30, 2026

Accepted May 26, 2026

Editor-in-Chief: Jessica E. Huber

Editor: Amanda I. Gillespie

https://doi.org/10.1044/2026_JSLHR-25-00923

ABSTRACT

Purpose: The purpose of this study was to develop and test standardized stimuli for the Pediatric Auditory-Perceptual Evaluation of Voice protocol. The use of these stimuli for cepstral/spectral voice measures was also considered.

Method: Three sets of standardized sentences and four prompts designed to elicit connected speech were developed and tested in order to identify the most effective stimuli for use with children. One hundred forty children with ($n = 50$) and without ($n = 90$) benign vocal fold lesions recorded all potential stimuli, as well as the sentences developed for the adult Consensus Auditory-Perceptual Evaluation of Voice protocol. Sentences were evaluated for production accuracy, and cepstral/spectral voice measures were calculated to determine whether potential stimuli would separate dysphonic and normophonic voices. Spontaneous speech prompts were assessed for duration of speech sample elicited. Subsequently, auditory-perceptual voice assessments were performed using the newly developed pediatric stimuli. Production accuracy, cepstral/spectral voice measures, and auditory-perceptual rating reliability were compared across age, sex, and voice diagnosis.

Results: Children produced pediatric sentences significantly more accurately than sentences developed for adults. The best produced sentence set resulted in over 80% accuracy across children 3 through 17 years of age. Cepstral peak prominence calculated on these sentences was significantly lower for children with voice disorders when compared to those with typical voices. Spontaneous speech prompts elicited an average of about 40 syllables of phonation over 25–30 s in children ≤ 6 years of age, and voice samples increased in duration for older children. Auditory-perceptual ratings of voice significantly separated children with and without voice disorders and produced reliability levels similar to those previously reported for adults.

Conclusion: Findings indicate that the developed pediatric stimuli are effective for auditory-perceptual assessments across childhood and can be produced more accurately than sentences designed for adults.

Voice disorders disrupt communication for an estimated 6.7% of school-aged children and 7.4% of adolescents in the United States (Fujiki & Thibeault, 2024a, 2024b, 2025a), although true prevalence is likely higher as many pediatric voice problems go unreported (Carding et al., 2006).

Broadly defined, voice disorders involve alterations in vocal quality, pitch, or loudness (all known as dysphonia) that fail to meet an individual's communication needs or attract negative attention considering their age and gender (Aronson & Bless, 2009). In the majority of pediatric cases, dysphonia is attributed to benign vocal fold lesions, such as vocal fold nodules, polyps, and cysts (Akif Kiliç et al., 2004; Campano et al., 2023; Nardone et al., 2014). The impact of pediatric voice disorders can be substantial (Ilić-Savić & Petrović-Lazić, 2023). Beyond disrupting communication, dysphonia is associated

Correspondence to Robert Brinton Fujiki: rfujiki@iu.edu. **Disclosure:** The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

with adverse emotional and social consequences (Connor et al., 2008; Fujiki & Thibeault, 2025b; Verduyck et al., 2011), including negative perceptions by peers and adults (Lass, Ruscello, Bradshaw, & Blankenship, 1991; Lass, Ruscello, Stout, & Hoffman, 1991; Ma & Yu, 2013).

Voice disorders are evaluated using laryngeal imaging, as well as a battery of acoustic, aerodynamic, and patient-reported measures (Patel et al., 2018). Auditory-perceptual voice assessments, however, most holistically quantify voice quality and measure treatment progress (Barsties & De Bodt, 2015; Feinstein & Abbott, 2021; Fujiki & Thibeault, 2021, 2023). Auditory-perceptual assessment occurs when a clinician (usually a speech-language pathologist [SLP]) listens to an individual's voice to evaluate and document deviations in overall voice quality as well as individual parameters of voice quality (i.e., breathiness, roughness, strain) with regard to an individual's age, sex, and gender. These assessments are particularly crucial for children, who may not tolerate the collection of quantitative voice measures (Brehm et al., 2009; Kelchner et al., 2012). For example, under 17% of 3- to 4-year-old children tolerate videostroboscopy, and even fewer successfully complete laryngeal aerodynamic assessments (Fujiki, Venkatraman, Godbout, & Thibeault, 2025). Only half of 3- to 4-year-olds produce sustained vowels with the consistency of loudness and pitch required for acoustic voice measures (Fujiki, Venkatraman, Godbout, & Thibeault, 2025). In fact, 40%–50% of older children (6–9 years of age) do not tolerate videostroboscopy, and 30%–40% cannot complete laryngeal aerodynamic measures (Fujiki, Venkatraman, Godbout, & Thibeault, 2025). Although many older children can complete acoustic assessments, not all are able to do so. Additionally, available normative data are limited, particularly given that norms must be age, sex, and gender specific for pediatric populations (Fujiki, Venkatraman, & Heller Murray, 2025; Maturo et al., 2012). As such, auditory-perceptual evaluations may represent a clinician's best opportunity to quantify and document a child's dysphonia.

Little is known, however, about how auditory-perceptual voice assessments function in children (Fujiki & Thibeault, 2021). Past studies examining pediatric auditory-perceptual voice assessments in children have utilized protocols designed for adult voices (Hartley et al., 2017), such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Braden & Thibeault, 2020; Fujiki & Thibeault, 2022). The CAPE-V was created to increase the reliability and validity of auditory-perceptual assessments through standardization of rating stimuli and rating scale (Kempster et al., 2009; Zraick et al., 2011). In this protocol, voice quality is rated based on a specific voice sample consisting of standardized sentences, a sustained vowel, and a spontaneous speech sample (Kempster et al., 2009). The CAPE-V uses visual analog scales (VASs) to rate vocal parameters such

as roughness, breathiness, strain, pitch, and loudness (Kempster et al., 2009). VASs may improve specificity and reliability of ratings and produce data that can be easily analyzed statistically (Fujiki & Thibeault, 2021; Hillman, 2003; Kempster et al., 2009; Nemr et al., 2012).

Reliable auditory-perceptual voice assessment using VAS may be feasible in pediatric populations (Feinstein et al., 2025); however, preliminary studies suggest that the adult CAPE-V protocol may not be well adapted for children (Fujiki & Thibeault, 2021; Kelchner et al., 2010; Masaki, 2001). As a major concern, the standardized CAPE-V stimuli includes six to nine syllable sentences designed to elicit specific laryngeal behaviors (Kempster et al., 2009). Standardized stimuli are important because the accuracy of auditory-perceptual assessments improves when ratings are made from consistent speech samples (Kreiman et al., 2007). Additionally, standardized stimuli can be used for acoustic analyses to assess dysphonia severity and track treatment progress (Fujiki et al., 2023; Watts, 2015). The CAPE-V protocol also includes sustained vowel production, but the reliability of auditory-perceptual voice ratings improves when ratings are made using connected speech (Costa et al., 2023; Law et al., 2012; Lechien et al., 2018; Maryn & Roy, 2012). The CAPE-V sentences have been adapted for numerous languages and cultures (Calaf & Garcia-Quintana, 2024; Mahalingam et al., 2024; Pommée et al., 2026), and past investigation suggests that these sentences can be used to efficiently assess parameters such as speaking fundamental frequency (F_0 ; Sandage et al., 2015) as well as overall severity of dysphonia (Fujiki et al., 2023).

Although the CAPE-V sentences have clinical utility for adults, young children may not be able to produce these sentences accurately due to linguistic, articulatory, and memory demands (Kelchner et al., 2010). This is unfortunate considering that extra phonemes, pauses, misspeaking, and/or abnormal prosody associated with high cognitive load may impact CAPE-V ratings and influence evaluation of F_0 or other acoustic parameters (Aoki et al., 2021; Demirci et al., 2021). For example, changes in pitch and loudness can impact acoustic analyses (Ataee et al., 2022; Brockmann et al., 2008; Gramming et al., 1988; Murton et al., 2020) and may well distract from auditory-perceptual assessment.

In addition to standardized stimuli, the CAPE-V protocol also includes eliciting a spontaneous speech sample. In order to elicit sufficient extent and phonation of connected speech to permit analysis (Barsties & Maryn, 2017; Englert et al., 2022), the CAPE-V provides conversational prompts (i.e., *Tell me about your voice problem*, *Tell me about your last vacation*). These specific prompts, however, often elicit little or no verbal response from children (Reynolds et al., 2025; Reynolds & Fleury, 2024).

In short, the tasks utilized in the CAPE-V are designed for use with adults, and children may respond quite differently to the prompts employed. As time is short and children have limited tolerance for evaluation procedures, clinicians would benefit from effective standardized stimuli to elicit adequate voicing for auditory-perceptual voice assessments. In developing and validating such stimuli, it would be important to determine that children can complete the required tasks. For example, children should be able to read or repeat stimulus sentences accurately without long pauses, misspeaking, or other injections (i.e., *uh*, *umm*). Sentences or phrases must be short enough for children to produce but long enough for SLPs to assess voice quality. In addition, to maximize clinical utility, acoustic measures taken on these sentences should separate dysphonic from normophonic voices. Spontaneous speech prompts should consistently elicit speech samples of a sufficient duration for clinicians to evaluate voice quality. Furthermore, for ease of use, the same stimuli should ideally be appropriate across childhood—for very young children through adolescents. Finally, and perhaps most importantly, stimuli for children should facilitate the production of reliable auditory-perceptual ratings of voice in both dysphonic and normophonic voices.

The focus of the current study was to develop and validate a standard set of stimuli for the Pediatric Auditory-Perceptual Evaluation of Voice (PAPE-V) protocol. Multiple sets of standardized sentences were designed for children. These sentences were tested against each other and compared with the original CAPE-V sentences designed for adults. Additionally, spontaneous speech prompts were presented to children to characterize the nature of the voice samples they elicit across childhood. It was hypothesized that children would produce PAPE-V sentences designed for pediatric populations more accurately than stimuli designed for adults. Additionally, it was hypothesized that cepstral/spectral measures of voice would separate dysphonic and normophonic voices when calculated on the PAPE-V sentences. Finally, auditory-perceptual voice assessments were performed on children using the PAPE-V stimuli, as well as the adult CAPE-V stimuli for a subset of children. It was hypothesized that auditory-perceptual voice assessments in children would be more reliable when performed using stimuli designed for use with children.

Method

Study Design

This study created and validated stimuli designed to elicit voice production in pediatric auditory-perceptual

voice evaluations. This study represents the first step in the development of the PAPE-V protocol. To this end, three sets of standardized sentences (to be read or repeated) and four prompts for connected speech were developed and tested. Children with and without voice disorders between 3 and 17 years of age were recruited to record these stimuli. Children were also recorded producing the stimuli created for the adult CAPE-V protocol (this study commenced prior to release of the revised CAPE-V [Kempster et al., 2025]). Spontaneous speech prompts were compared based on the duration of speech samples elicited. Sentences underwent three stages of evaluation to determine which stimuli were most effective for use with pediatric populations. First, accuracy of sentence production was assessed across age groups. Second, the most accurately produced stimuli were analyzed using cepstral/spectral measures of voice. Voice measures were compared across age, sex, and voice status. Finally, auditory-perceptual analyses were performed using the newly developed PAPE-V stimuli in order to ensure that reliable voice ratings could be produced with them. Auditory-perceptual ratings were made from recordings using VASs (protocol described further below). For a subset of children ≥ 7 years old, auditory-perceptual ratings were also performed on the stimuli developed for the original CAPE-V protocol to allow for comparison across adult and pediatric stimuli. All study procedures were approved by the University of Wisconsin–Madison Institutional Review Board (#2023-0227).

Participants

Children

Children with and without diagnosed voice disorders, 3–17 years of age, were recruited for this study. Children were recruited from an outpatient voice and swallow clinic associated with a tertiary children's hospital, as well as from the surrounding community. Children presenting to the clinic for non-voice-related complaints were considered eligible for participation if they met all other inclusion criteria. Inclusion criteria dictated that children were (a) 3–17 years of age, (b) native English speakers, and (c) either had been diagnosed with a voice disorder or had been denied any history of voice complaints. Children diagnosed with voice disorders underwent videostroboscopy (strobe light) or laryngoscopy (steady-state halogen light) and were diagnosed with laryngeal pathology by a pediatric otolaryngologist and SLP. Children without voice disorders did not undergo laryngeal imaging but were screened for any history of voice complaints using both self- and parent report. Exclusionary criteria for all children included (a) neurologic conditions preventing consent, (b) head or neck cancer, (c) dysarthria of speech, or (d) diagnosed respiratory conditions, which would

impact phonation. Children were divided into three age groups, ≤ 6 years, 7–10 years, and 11–17 years of age. Caregivers gave informed consent, and children were assented prior to participation.

SLP Raters

Voice-specialized SLPs were recruited from across the United States. SLPs were identified by contacting clinicians from pediatric voice clinics or research laboratories throughout the country. All SLPs were required to have a minimum of 5 years' experience treating voice disorders in clinical or research settings. Additionally, SLPs were required to be (a) professionally certified SLPs (possessed current certificates of clinical competence), (b) native English speakers, and (c) specialized in the treatment of voice, resonance, and upper airway disorders in children and adolescents. Finally, SLPs were also required to have access to a computer with internet capabilities and a private room where they could listen to voice samples. SLPs were currently practicing in California, Iowa, Massachusetts, Ohio, Pennsylvania, and Wisconsin. SLPs were compensated for study participation.

Stimuli Development

Sentences were developed by the research team to match the phonemic targets set for the original CAPE-V sentences (Kempster et al., 2009), but were shortened and simplified. Sentences and supporting rationale are presented in Table 1. Sentences featured the following sounds: (a) coarticulation of /a, i, u/, (b) soft glottal attacks (as well as voiceless to voiced transitions), (c) all voiced phonemes, (d) hard glottal attacks, (e) nasal consonants, and (f) oral pressure sounds. Sentence length was limited to three to four syllables in most cases; however, five to six syllables were permitted if needed to improve sentence clarity. Sentences were designed to be within reading level for children ≥ 7 years of age. The three- to

four-syllable target was selected to ensure sentences were long enough to produce voice samples but short enough to be accurately produced by young children. References to culture-specific holidays (i.e., Easter) and "hitting" were removed to promote inclusivity and avoid encouraging undesirable behaviors. Sentences were informally piloted with pediatric voice patients for several months prior to study initiation to ensure reproducibility. A detailed analysis of sentences is presented in the Results section.

Children also produced four speech samples using various prompts. Elicitation order was counterbalanced across participants. Elicitation prompts were as follows: (a) "Tell me about your family," (b) "Tell me about your favorite movie," and (c) "Tell me how to make a peanut butter and jelly sandwich." For the fourth prompt, children were asked to describe the updated Cookie Theft picture from the Boston Diagnostic Aphasia Examination ("Describe what you see in this picture"; Roth, 2011). Prompts were selected based upon current clinical practices and past research in adult populations (Gilman et al., 2023).

Voice Recordings

Recordings were made at a 44.1-kHz sampling rate (24 bits) using the Computerized Speech Lab (CSL) by PENTAX (mouth-to-mic distance of 2 in.; Patel et al., 2018). Gain staging ensured adequate signal strength and avoided peak clipping. Sentences were rerecorded if peak clipping was observed on the CSL interface. Participants were asked to read or repeat the four sets of six sentences created to elicit specific laryngeal behaviors (order counterbalanced). For children in the ≥ 6 years old group, the SLP verbally modeled each sentence immediately prior to elicitation and recording of the participant's sample. Children in the 7–10 years old group were allowed to choose whether they wanted to read or repeat the sentences, and children in the 11–17 years old group were asked to read the sentences unless they felt uncomfortable doing so. When children

Table 1. Adult and pediatric sentences and rationale.

Sentence type	Adult sentences	Pediatric sentences		
Rationale	Original adult CAPE-V sentence	Chosen sentences	Other sentences	
Coarticulation of /a, i, u/	The blue spot is on the key again.	You and Bob are eating.	He gets the big balloon.	The dog is brown.
Soft glottal attacks (voiced to voiceless transitions)	How hard did he hit him?	Harry has a hat.	He has honey.	He has hair.
All voiced	We were away a year ago.	We were away.	We want wool.	Wally was wet.
Hard glottal attacks	We eat eggs every easter.	We eat eggs.	Every elephant eats.	Eyes and ears.
Nasal consonants	My mama makes lemon muffins.	Mama made muffins.	No more monkeys.	Make more money.
Oral pressure sounds	Peter will keep at the peak.	Pet the puppy.	Pick a pretty poppy.	Pass the popcorn.

Note. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) sentences from Kempster et al. (2009).

chose to read, sentences were presented on a printed sheet. A sustained vowel (/a/ for 5 s) was also collected.

Conversational prompts were tested in counterbalanced order. After each prompt was presented, children were allowed to speak as long as they chose. When children responded to the prompt, SLPs limited their reactions to facial expressions with as little phonation as possible. Single word responses such as “Wow,” “Oh,” and “Why?” were produced to keep children engaged. If children declined to speak, or produced very limited responses to a prompt, then basic follow-up questions were utilized (i.e., “Tell me more”). Additionally, prompt-specific follow-up questions were sometimes used if a child did not respond. For example, the prompt “Tell me about your family” might be followed up by “Who is oldest in your family?” in order to get a child speaking. Participants were asked to produce all stimuli at comfortable pitch and loudness.

Rating Procedures

SLP raters were presented with 73 unique voices (order randomized). A random number generator was used to determine which recordings would be presented to SLP raters. Seventy-three voices were rated in order to have enough voices to calculate reliability, but few enough to avoid major fatigue effects for raters (based on pilot testing). For each voice, SLPs were presented with the six standardized sentences, a sustained vowel (/a/), and a conversational speech sample. Conversational speech samples were cut to include no more than 20 s of connected speech in order to mimic clinical situations where a child may only produce limited voicing. SLP raters were blinded to child voice status (voice disorder vs. normophonic), as well as any other demographic information. As auditory-perceptual voice ratings, particularly of pitch and loudness, are context dependent, SLPs were told the gender and age of each child. Voice samples presented to raters were balanced across age, sex, and voice status (normophonic vs. voice disorder). Fifteen percent of voices were rated twice to allow for the calculation of intrarater reliability ($n = 11$). Repeated voices were also chosen using a random number generator. Additionally, 30 voices of children ≥ 7 years of age were presented a second time using the adult CAPE-V sentences (order counterbalanced between children). Accuracy on adult sentences was too low for children < 7 years of age. Power analysis (using G*Power) for repeated-measures analysis of variance ($f = .3$, $\alpha = .05$, $\beta = .95$) indicated that 30 voices would need to be rated using both adult and pediatric sentences to identify potential differences between stimuli. Audio files were shared with raters via secure Box.com cloud storage in uncompressed format (i.e., .wav format files) to ensure recording fidelity. Raters were asked to listen to samples at consistent volume, using research-grade headphones in a

quiet room. Headphones were shipped to SLP raters prior to study commencement (Pro-10, Universal Product No. 709466429447, OneOdio).

Auditory-perceptual ratings were performed using VAS anchored with the terms *normal* (0) on the left and *severely disordered* (100) on the right. Raters placed slider buttons on VAS at the point that they felt corresponded with the severity of that particular voice parameter. Possible values included 0 through 100 with all possible integers. As SLPs moved the slider button, they were able to see the corresponding numerical value. SLPs rated voices for overall voice quality, roughness, breathiness, strain, pitch, and loudness. If raters indicated that pitch or loudness were not normal (scored anything higher than 0), they were asked whether these parameters were “too low” or “too high.” One set of ratings was made from standardized sentences, sustained vowel, and spontaneous speech sample. Voice ratings were submitted online using Qualtrics.

In addition to rating voice quality, SLPs were also asked to rate the quality of the recording for each voice. Raters responded to a multiple-choice question (“Did anything about the recording make it difficult to rate this voice?”) by indicating whether recording quality, sample duration, loudness, or any other parameter hindered their ability to rate the recording (answer options: [a] no problem, [b] low vocal volume, [c] poor recording quality, [d] background noise, [e] short voice sample, [f] other—with a fill-in-the-blank option). If more than 50% of raters indicated that the recording quality or background noises were problematic, the sample was not used in analysis. As variations in loudness and sample duration are typical aspects of pediatric voice assessments, voice recordings were still considered even if these options were selected. No recordings met the threshold for exclusion based on SLP ratings of recording quality.

Outcome Measures

Sentences: Production Accuracy

First, a researcher rated whether each sentence was produced accurately using a forced-choice rating system (accurate or not). Sentences were considered inaccurate if any of the following errors were observed: (a) disfluencies such as repetitions or blocks, (b) mispronunciations of words, (c) misreading any word in the sentence, or (d) pauses longer than 1.5 s in the middle of a sentence. Praat was used to visualize the speech signal and assess the duration of pauses. Disfluencies, mispronunciations, or incorrect productions were assessed by listening to the audio output. Twenty percent of sentences were rated by a second rater to allow for calculation of inter- and intrarater reliability.

Sentences: Acoustic Analysis

Cepstral/spectral measures of voice were used to determine if stimuli separated normophonic and dysphonic children. Cepstral peak prominence (CPP), CPP standard deviation (CPPSD), low-to-high ratio (LHR), and LHR standard deviation (LHRSD) were calculated. CPP, a measure of peak cepstral amplitude (Gómez-García et al., 2019; Heman-Ackah et al., 2003), was chosen because it is sensitive to dysphonia in children (Awan et al., 2009; Fraile et al., 2013; Watts & Awan, 2011). Additionally, LHR, an estimation of spectral tilt above and below 4 kHz, effectively supplements CPP for characterizing dysphonic voices (Gómez-García et al., 2019). Initial research indicates that these measures can be calculated on pediatric voices and are sensitive to the improvements in voice quality produced by voice therapy (Fujiki et al., 2023). Acoustic samples were analyzed using the all-voiced sentence protocol in the Multi-Dimensional Voice Program Analysis of Dysphonia (PENTAX) in Speech and Voice. Speech samples were analyzed even if sentences contained minor production errors (pauses, misspeaking, etc.); however, if major errors were present (i.e., incomplete productions), the child's second attempt at a sentence was used for acoustic analysis. Inter- and intrarater reliability were calculated on 15% of acoustic analyses to ensure accuracy.

Spontaneous Speech Samples

Conversational speech samples were analyzed for number of utterances and duration in seconds to determine which prompts produced the longest speech samples for each age group. As an efficient prompt should produce as much phonation as possible as quickly as possible, prompt effectiveness was measured as number of utterances divided by total duration of the sample in seconds.

Auditory-Perceptual Assessments

Variability analysis (or the square of the difference between ratings) was used to assess both inter- and intrarater reliability across various populations/conditions. Raw VAS ratings were also considered to determine whether values differed across those with and without voice disorders. Additionally, intraclass correlation coefficients (ICCs) were calculated for the overall sample in order to allow for comparison with ICC ranges reported in past studies of auditory-perceptual voice assessments.

Statistical Analysis

Descriptive statistics were calculated for all variables. Accuracy of sentence production was calculated as a percentage of children who produced each sentence correctly. Binary logistic regression was used to compare sentence accuracy across age group (≤ 6 , 7–10, 11–17 years),

sex (male, female), sentence type (six sentence types), and voice status (voice disorder, typical).

A multivariate linear mixed-effects model was used to compare cepstral/spectral measures across children with and without voice disorders. CPP, CPPSD, LHR, and LHRSD were modeled jointly in a multivariate framework to account for correlations among these measures. Data met necessary assumptions or were appropriately transformed. Fixed effects included sex (male, female), age group (≤ 6 , 7–10, 11–17 years), and voice status (voice disorder, typical). Because each participant produced all six sentence types, sentence type was treated as a repeated measure with random intercepts for participant. Additionally, all interactions between sex, age, and voice status were included in the model. Interactions with sentence type were initially probed but were not significant, and model fit improved when these were removed. Tukey-adjusted post hoc comparisons were used as appropriate. A separate multivariate linear mixed-effects model was used to compare cepstral/spectral values across the pediatric sentences and those from the adult CAPE-V protocol. Fixed effects included age (7–10, 11–17 years), sex, and voice status. Participant was included as a random effect to account for the repeated-measures nature of the study design.

For inter- and intrarater reliability, variability analysis, or the square of the difference between ratings, was used to demonstrate the extent to which a rating (or rater) differed from other ratings (or raters; Chan & Yiu, 2002; Dahl et al., 2021; Eadie & Kapsner-Smith, 2011). Lower numbers indicated better agreement. Correlations between repeated ratings were computed and averaged across voice samples to quantify intrarater reliability (i.e., the consistency of raters). Variability scores for ratings made with the pediatric stimuli were compared using multivariate models. Models included sex (male, female), voice status (voice disorder, typical), and a Sex $\times$ Voice Status interaction. Age was initially included in the model but was not significant and was therefore removed to improve model fit. Tukey's honestly significant difference tests were used for post hoc comparisons. A multivariate generalized linear mixed model was used to compare the variability of auditory-perceptual ratings made using adult and pediatric sentences. Fixed effects included age (7–10 and 11–17 years), sex, and voice status. Participant was included as a random factor.

As overall interrater reliability has often been assessed using ICCs (two-way random model), ICCs for the overall study population were calculated to facilitate comparison with past work. Additionally, ICCs were calculated for all ratings made using the adult CAPE-V sentences. Across all analyses, α was set at .01 in order to protect

against Type I error. All statistical analyses were performed in SPSS Version 30.

Results

One hundred forty children were recruited to record stimuli (males participants = 67, $M_{\text{age}} = 9.0$, $SD = 1.2$; female participants = 73, $M_{\text{age}} = 9.3$, $SD = 1.2$). Fifty of these children had been diagnosed with benign vocal fold lesions, and 90 denied any history of voice problems. Counts and ages for children with and without voice disorders are reported in Table 2.

Reliability

For measurement of sentence production accuracy, both inter- and intrarater reliability were in excellent range with an ICC of .98 (95% confidence interval [CI] [.96, .99]). Similarly, inter- and intrarater reliabilities were in excellent range for analysis of spontaneous speech prompts with ICCs of .97 (95% CI [.94, .99]) and .96 (95% CI [.93, .98]), respectively. For analysis of cepstral/spectral measures, interrater (ICC = .95, 95% CI [.92, .97]) and intrarater reliabilities (ICC = .96, $p < .001$, 95% CI [.93, .98]) were also in excellent range.

Sentence Production Accuracy

Production accuracy across all sentences and age groups is presented in Table 3. Across all sentence types, age group significantly predicted sentence accuracy as children in the 7–10 years old age group were twice as likely (odds ratio [OR] = 2.1, $SE = .28$, $p < .001$) and children

in the 11–17 years old group over 11 times as likely ($OR = 11.4$, $SE = .46$, $p < .001$) to produce sentences accurately when compared with children ≤ 6 years of age. Accuracy ranged from 45% to 97% for children 11–17 years of age but dropped as low as 3% for children ≤ 6 years of age, and as low as 36% for children 7–10 years of age.

Additionally, sentence type predicted accuracy, as children were 66% less likely to produce the original CAPE-V sentences correctly when compared to the new stimuli ($OR = .34$, $SE = .34$, $p < .001$). No other significant differences were observed. For children ≤ 6 years of age, accuracy on the original CAPE-V sentences was poor across all sentence types. Accuracy ranged from 39% for “the blue spot is on the key again” to 3% for “Peter will keep at the peak.” For children in the 7–10 years old age group, accuracy was again lowest on “Peter will keep at the peak” (36%) and ranged between 61% and 72% for all other sentence types. For children between 11 and 17 years of age, accuracy on the original sentences was high ($> 87\%$) for all sentences, except “Peter will keep at the peak,” where accuracy dropped to 45%. Inaccuracies were largely caused by misspeaking or substituting alternate words (63%), incomplete productions (28%), or long pauses (9%). Regarding the newly developed sentences, the best accuracy for children ≤ 6 years of age ranged from 80% to 94% across sentence types. For children 7–10 years of age, the best accuracy ranged from 83% to 90% across sentence types, and for children 11–17 years of age, accuracy was between 93% and 97%. Neither voice status nor sex were significant predictors of production accuracy.

Sentence Selection

Significant differences were not observed between any of the newly developed pediatric sentences. In most

Table 2. Cepstral/spectral measure means and standard deviations across age and sex for children with and without voice disorders.

Voice status	Sex	Age group, years	<i>n</i>	<i>M</i> _{age}	CPP		CPPSD		LHR		LHRSD	
					<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Typical voices	Female	< 6	15	5	5.8	1.0	2.9	0.6	24.7	5.1	7.6	2.6
		7–10	15	8.9	6.0	1.5	2.9	0.9	26.9	4.6	8.2	2.4
		11–17	15	13.7	8.8	1.2	2.5	0.7	30.5	3.6	6.8	2.2
	Male	< 6	15	4.9	5.9	1.0	2.9	0.6	25.9	4.6	7.7	3.0
		7–10	15	7.8	6.6	1.6	2.9	0.9	26.3	4.9	6.9	1.6
		11–17	15	14.6	11.2	1.8	2.5	0.7	31.4	3.1	6.9	1.8
Voice disorders	Female	< 6	8	4.8	4.2	1.9	2.9	0.7	20.2	2.3	7.9	2.2
		7–10	9	9	4.5	1.6	2.7	0.5	24.4	5.4	7.9	2.9
		11–17	8	14.8	5.4	1.3	2.6	0.7	25.4	3.1	6.4	2.4
	Male	< 6	8	5.1	5.0	1.9	2.7	0.6	23.1	5.3	8.3	2.6
		7–10	9	8.1	4.4	1.7	2.5	0.7	23.7	5.8	8.4	2.6
		11–17	8	13.6	6.3	1.1	3.2	1.3	27.6	4.8	7.0	2.2

Note. CPP = cepstral peak prominence; CPPSD = CPP standard deviation; LHR = low-to-high ratio; LHRSD = LHR standard deviation.

Table 3. Sentence production accuracy across age groups and sentence types.

Sentence type	Coarticulation of /a, i, u/	Soft glottal attacks	All voiced	Hard glottal attacks	Nasal consonants	Oral pressure sounds
Age groups, years	You and bob are eating.	Harry has a hat.	We were away.	I eat eggs.	Mama made muffins.	Pet the puppy.
< 6	86%	92%	94%	92%	80%	92%
7–10	89%	86%	89%	90%	83%	90%
11–17	97%	93%	97%	94%	97%	97%
Age groups, years	He gets the big balloon.	He has honey.	We want wool.	Every elephant eats.	No more monkeys.	Pick a pretty poppy.
< 6	72%	78%	56%	78%	77%	53%
7–10	77%	78%	75%	75%	86%	78%
11–17	97%	97%	90%	93%	97%	93%
Age groups, years	The dog is brown.	He has hair.	Wally was wet.	Eyes and ears.	Make more money.	Pass the popcorn.
< 6	67%	69%	64%	67%	56%	72%
7–10	86%	83%	81%	81%	83%	81%
11–17	97%	93%	90%	93%	76%	97%
Age groups, years	The blue spot is on the key again.	How hard did he hit him?	We were away a year ago.	We eat eggs every easter.	My mama makes lemon Jam.	Peter will keep at the peak.
< 6	39%	25%	36%	30%	31%	3%
7–10	64%	61%	72%	61%	64%	36%
11–17	90%	87%	90%	90%	87%	45%

Note. Consensus Auditory-Perceptual Evaluation of Voice sentences from Kempster et al. (2009).

cases, however, differences of over 10% were observed between the most accurately produced sentence and the other alternatives (see Table 3). These differences were largely observed in children ≤ 6 years of age and those 7–10 years of age. Thus, in most cases, the most accurate sentence across all age groups was selected. For the nasal consonant sentence, however, two sentences were produced with similar accuracy. Specifically, “mama made muffins” and “no more monkeys” were both produced with roughly 80% accuracy by young children. As such, “mama made muffins” was selected because it more closely aligns with the adult CAPE-V stimuli and would likely be easier for clinicians who need to transition to the adult sentences as children age. Final sentence selection is presented in Table 1.

Spontaneous Speech Prompts

Syllables

A significant effect of age was observed for number of syllables produced, $F(2, 554) = 18.3$, $p < .001$. Children in the 11–17 years of age group produced significantly more syllables than children in the 7–10 years of age group ($p < .01$) and the ≤ 6 years of age group ($p < .01$). Similarly, children in the 7–10 years of age group produced significantly more syllables than those in the ≤ 6 years of age group ($p < .01$). Means for number of syllables produced across age and prompt are presented in Table 4.

Differences between prompts varied across age groups, as evidenced by a significant interaction between age group and prompt, $F(6, 544) = 4.6$, $p < .001$. Prompts asking the child to describe how to make a sandwich or describe their family produced the greatest number of syllables across all age groups. For children ≤ 6 years of age, however, describing the Cookie Theft picture produced significantly fewer syllables than all other spontaneous speech prompts ($p < .005$). For children 7–10 years of age, describing the Cookie Theft picture produced significantly fewer syllables than asking about the child’s family or describing how to make a sandwich ($p < .001$). For this group, no significant differences were observed between the number of syllables elicited in response to the Cookie Theft picture and asking about the child’s favorite movie. For children 11–17 years of age, no significant differences between prompts were observed.

Duration in Seconds

For speech duration in seconds, a significant main effect of prompt was observed, $F(3, 544) = 18.5$, $p < .001$. The longest speech samples were produced by asking the child about their family or how to make a sandwich. These prompts produced significantly longer speech samples than asking children about their favorite movie ($p < .001$) or describing the Cookie Theft picture ($p < .001$). No significant effect of age ($p = .082$) or interaction ($p = .342$) were observed. Means are presented in Table 4.

Table 4. Means and standard deviations for spontaneous speech sample measures.

Age group, years	Measure	Describe the Cookie Theft picture.		Tell me about your family.		What is your favorite movie?		Tell me how to make a peanut butter and jelly sandwich.	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
< 6	Syllables	21.6	6.9	40.4	18.3	33.5	25.4	41.2	17.1
	Seconds	31.1	14.0	25.8	9.4	19.8	12.8	27.0	9.1
	Syllables per second	0.9	0.6	1.6	0.6	1.7	0.6	1.5	0.5
7–10	Syllables	29.9	16.5	56.7	19.7	38.7	29.8	60.5	19.1
	Seconds	31.4	16.2	28.3	6.7	23.3	15.7	30.4	7.1
	Syllables per second	1.1	0.6	2.0	0.4	1.6	0.5	2.0	0.4
11–17	Syllables	37.9	17.6	39.1	19.7	34.1	18.3	46.9	19.7
	Seconds	38.0	18.2	26.4	16.2	22.0	13.8	29.3	15.7
	Syllables per second	1.2	0.6	1.6	0.5	1.6	0.4	1.8	0.5
Total	Syllables	29.8	15.9	45.7	20.8	35.5	25.0	49.8	20.3
	Seconds	33.4	16.4	26.9	11.3	21.7	14.2	29.0	11.1
	Syllables per second	1.1	0.6	1.7	0.5	1.6	0.5	1.8	0.5

Syllables Produced per Second

A significant effect of age was observed for syllables produced per second, $F(2, 554) = 10.3$, $p < .001$, as children in the 11–17 years of age group produced significantly more syllables per second than children in the 7–10 years of age group ($p < .01$) or the ≤ 6 years of age group ($p < .01$). Children in the 7–10 years of age group produced more syllables per second than those in the ≤ 6 years of age group ($p < .01$). Means are presented in Table 4.

A significant interaction between age group and prompt was also observed for number of syllables produced per second, $F(6, 544) = 3.07$, $p = .006$. Patterns followed those observed for total number of syllables. Prompts asking the child to describe how to make a sandwich or describe their family produced the greatest number of syllables per second across all age groups. For children ≤ 6 years of age, however, describing the Cookie Theft picture produced significantly fewer syllables per second than all other spontaneous speech prompts ($p < .001$). For children 7–10 years of age, describing the Cookie Theft picture produced significantly fewer syllables per second than asking about the child's family or describing how to make a sandwich ($p < .001$). For this group, no significant differences were observed between the number of syllables produced by the Cookie Theft picture and asking about the child's favorite movie. For children 11–17 years of age, the only significant difference was that significantly more syllables per second were produced when children described how to make a sandwich than when children described the Cookie Theft picture ($p < .001$).

Prompt Selection

Based on these analyses, it was determined that the greatest amount of voicing was most efficiently elicited

when children were asked to describe their families or when asked to describe how to make a peanut butter and jelly sandwich. As such, it was determined that either prompt could reasonably be used.

Cepstral/Spectral Measures

CPP

Means are presented in Table 2. For CPP, a significant three-way interaction between voice status, age group, and sex was observed, $F(2, 540) = 6.61$, $p < .001$. In normophonic children, CPP increased with age. When compared to children ≤ 6 years of age, values were significantly greater in children 7–10 years of age for both male ($p < .001$) and female ($p = .01$) participants. Values significantly increased again for the 11–17 years of age group ($p < .01$); however, the magnitude of this change was significantly larger for male than female participants ($p < .01$; mean difference female participants = 2.8 dB, male participants = 4.6 dB). In children with voice disorders, CPP was in statistically similar range for children ≤ 6 years of age and those 7–10 years of age in both male and female participants. For female participants, values for the 11–17 years of age group were significantly greater than those for children ≤ 6 years of age ($p < .001$); however, CPP was statistically similar to the 7–10 years of age group ($p = .026$). In male participants, CPP for children 11–17 years of age was significantly greater than both the ≤ 6 ($p < .001$) and 7–10 ($p < .001$) years of age groups. The magnitude of this change was smaller than that observed in normophonic children (mean difference, male participants = 1.9 dB, female participants = .9 dB).

A significant Sentence Type (i.e., all voiced, voiceless) $\times$ Sex interaction was also observed for CPP,

$F(2, 540) = 5.73, p < .001$ (see Figure 1). CPP was significantly greater for male participants across the sentences “We were away,” “You and Bob are eating,” and “Pet the puppy” ($p < .001$). Sex differences were not significant across the other sentences.

Additionally, a significant effect of sentence type was observed, $F(5, 540) = 9.78, p < .001$. Post hoc analyses revealed that CPP values were significantly smaller for “Harry has a hat” when compared to all other sentences. All other sentences were in statistically similar range. Effects of age were identical across sentence types (see Table 5), as values were significantly greater for children in the 11–17 years of age group when compared with younger children (see Figure 1). Additionally, values were significantly higher for children with typical voices when compared to those with voice disorders across all sentence types (see Figure 2).

CPPSD

Means are presented in Tables 2 and 5. No significant effects of sex, voice status, age, sentence type, or significant interactions were observed for CPPSD.

LHR

A significant three-way interaction was observed between voice status, age group, and sex (see Table 2). This occurred because, for female participants with typical voices, LHR values increased with age. Values were significantly larger for children 7–10 years of age when compared to those ≤ 6 years of age ($p < .01$), and significantly larger for children 11–17 years of age when compared to those 7–10 years of age ($p < .01$). For typical male voices, LHR was in similar range for those ≤ 6 years of age and those 7–10 years of age. Values were significantly larger for children 11–17 years of age when compared to all other children ($p < .01$). For female participants with voice disorders, the largest differences in LHR were observed between children ≤ 6 years of age and those 7–10 years of age ($p < .01$). No significant differences were observed between LHR in female participants 7–10 years of age or those 11–17 years of age ($p = .03$). For male participants with voice disorders, the only significant difference was that children 11–17 years of age presented with significantly greater values than younger children ($p < .01$). This difference was smaller in magnitude than for male participants with typical voices.

LHRSD

For children with voice disorders, LHRSD values were significantly lower for male and female participants in the 11–17 years of age group when compared to children in the younger groups ($p < .01$). Female participants

with typical voices followed a similar pattern, with values being significantly smaller for children 11–17 years of age when compared with younger children. For male participants with typical voices, however, values were significantly lower for children 7–10 years of age ($p < .01$) and 11–17 years of age ($p < .01$) when compared to those ≤ 6 years of age. Means are presented in Tables 2 and 5.

Cepstral/Spectral Comparisons of Pediatric and Adult Sentences

No significant differences were observed between pediatric and adult sentences on any of the cepstral/spectral measures examined. Comparisons were only calculated for the two older age groups, as children ≤ 6 years of age did not produce these sentences with sufficient accuracy. Means are presented in Table 6.

Auditory-Perceptual Evaluations

Mean auditory-perceptual ratings are presented in Table 7. Across all parameters, values were significantly smaller for children with typical voices when compared to those with voice disorders. No significant effects of age or sex were observed on raw ratings.

ICCs

ICC values for auditory-perceptual ratings performed using the pediatric stimuli, as well as for the adult CAPE-V sentences, are presented in Table 8. Overall, ICCs for the pediatric stimuli ranged from .65 (overall voice quality) to .35 (for loudness). When using the adult sentences, ICC values ranged from .59 (overall voice quality) to .32 (loudness and pitch).

Inter- and Intrarater Variability Analysis

Means for variability data are presented in Table 9. For overall voice quality, a significant effect of voice status was observed, $F(1, 712) = 9.34, p = .002$, because variability scores were significantly higher for children with typical voices ($M = 620.6, SD = 1,050.9$) when compared to those with voice disorders ($M = 467.8, SD = 667.6$). Variability scores were higher for male than female participants; however, this difference was not statistically significant, $F(1, 712) = 5.02, p = .025$.

For breathiness, $F(1, 716) = 32.3, p < .001$, and roughness, $F(1, 716) = 12.6, p < .001$, variability scores were significantly larger for children with voice disorders than for those with typical voices. No significant effects of sex were observed. For strain, no significant effects of sex or voice status were observed.

Figure 1. Cepstral peak prominence (CPP) means and standard deviations across age, sex, and sentence type. *Indicates statistical significance.

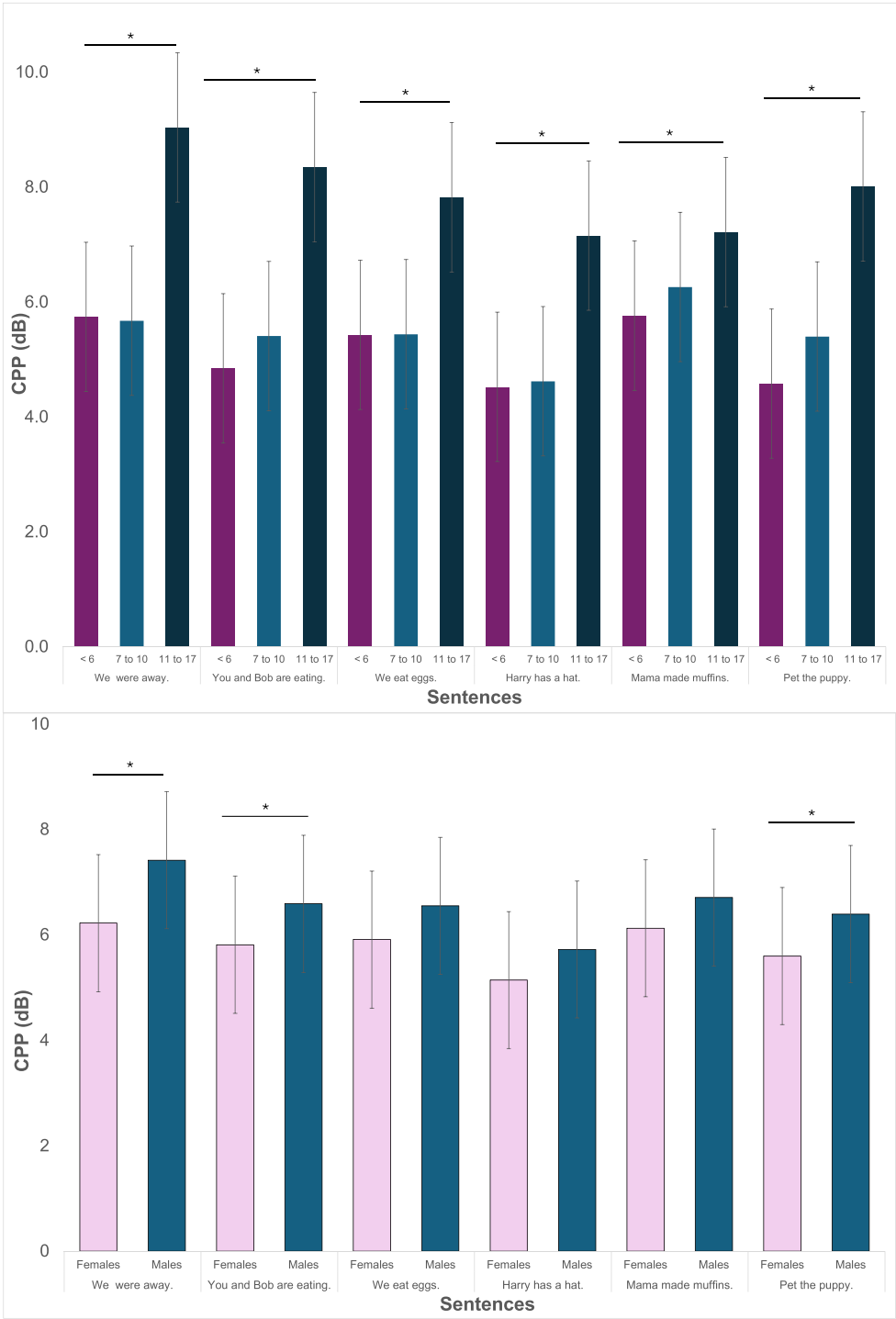


Table 5. Means and standard deviations for cepstral/spectral measures across sentence type and age.

Sentence	Age group, years	CPP		CPPSD		LHR		LHRSD	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
We were away.	≤ 6	5.7	1.0	2.9	0.6	27.5	5.1	6.9	2.6
	7–10	5.7	1.5	2.5	0.9	26.9	4.6	7.9	2.4
	11–17	9.0	1.2	2.4	0.7	29.9	3.6	6.4	2.2
You and Bob are eating.	≤ 6	4.8	1.8	2.6	0.7	24.2	5.0	7.0	2.5
	7–10	5.4	1.0	2.7	0.6	26.6	4.6	6.8	3.0
	11–17	8.3	1.6	2.9	0.7	28.6	4.9	6.6	1.6
We eat eggs.	≤ 6	5.4	1.8	3.1	0.7	19.8	3.1	8.7	1.8
	7–10	5.4	2.8	2.7	0.7	24.8	4.9	8.3	2.3
	11–17	7.8	1.0	3.0	0.6	28.1	4.9	6.6	2.8
Harry has a hat.	≤ 6	4.5	1.5	3.0	0.8	21.6	4.7	7.5	2.2
	7–10	4.6	1.9	2.9	0.7	23.3	3.4	7.1	2.0
	11–17	7.2	2.4	2.8	0.7	27.6	5.0	7.3	2.4
Mama made muffins.	≤ 6	5.8	1.9	2.9	0.7	25.0	2.3	9.6	2.2
	7–10	6.3	1.6	2.8	0.5	26.0	5.4	9.8	2.9
	11–17	7.2	1.3	2.8	0.7	28.6	3.1	7.3	2.4
Pet the puppy.	≤ 6	4.6	1.7	2.7	0.7	22.8	4.1	7.5	2.5
	7–10	5.4	1.9	2.5	0.6	24.2	5.3	7.3	2.6
	11–17	8.0	1.7	2.3	0.7	29.4	5.8	6.4	2.6

Note. CPP = cepstral peak prominence; CPPSD = CPP standard deviation; LHR = low-to-high ratio; LHRSD = LHR standard deviation.

Figure 2. Cepstral peak prominence (CPP) means and standard deviations across sentence type for those with and without voice disorders. *Indicates statistical significance.

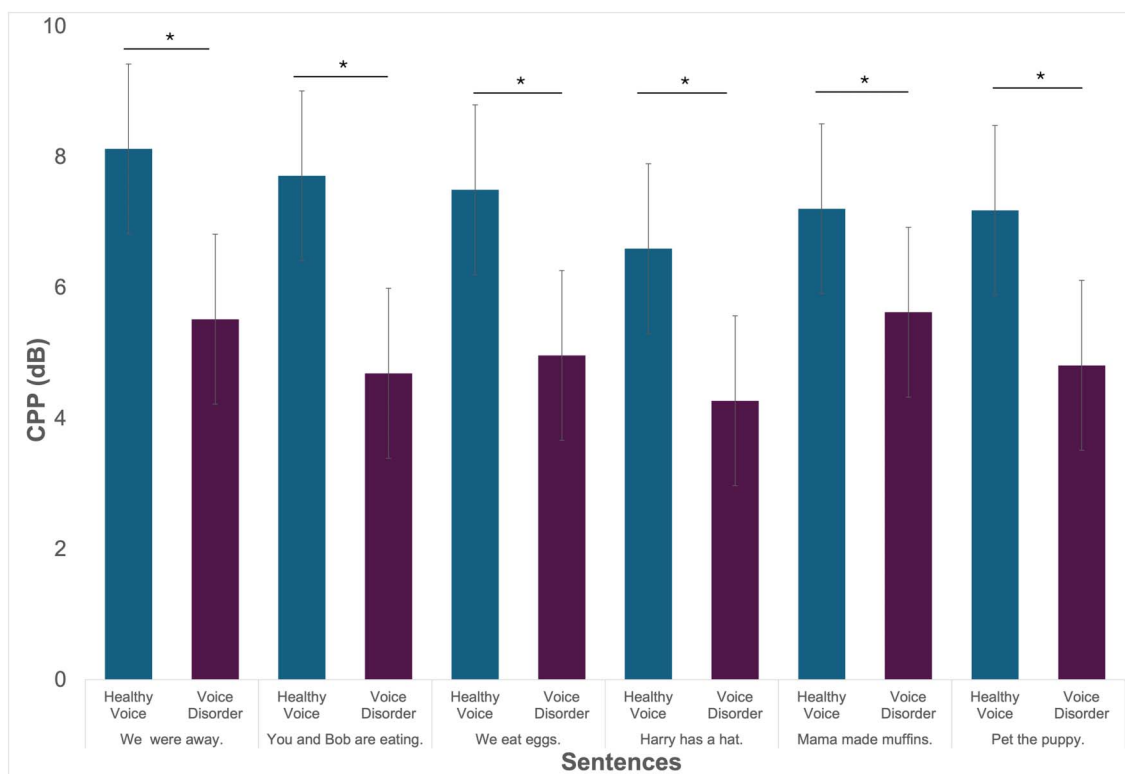


Table 6. Cepstral/spectral measures across pediatric and adult sentences.

Age group, years	Measures	Pediatric sentences		Adult sentences		<i>p</i> value
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Cepstral/spectral measures						
7–10	CPP	5.4	1.8	5.3	1.9	.56
	CPPSD	2.7	0.8	3.1	0.9	.65
	LHR	25.4	5.4	24.7	6.2	.78
	LHRSD	8.0	2.4	9.1	2.6	.32
11–17	CPP	8.6	2.6	7.9	2.9	.26
	CPPSD	2.7	0.9	3.1	0.5	.46
	LHR	29.4	4.2	27.5	6.3	.65
	LHRSD	6.8	2.1	7.2	2.6	.32

Note. CPP = cepstral peak prominence; CPPSD = CPP standard deviation; LHR = low-to-high ratio; LHRSD = LHR standard deviation.

Adult Versus Pediatric Stimuli

No significant differences in variability were observed across ratings made using adult and pediatric stimuli; however, variability scores were higher for ratings made using the adult CAPE-V sentences. Means are presented in Table 8.

Discussion

Auditory-perceptual voice assessments provide an efficient, effective method for holistically evaluating voice quality from a functional perspective (Kreiman et al., 1993). Since quantitative assessments are often impractical for children, auditory-perceptual evaluations become particularly important in identifying voice disorders (Deborah & Samayan, 2026; Gramuglia et al., 2014) and in monitoring progress in voice therapy (Fujiki et al., 2023). Research indicates that auditory-perceptual ratings are most reliable when performed on standardized stimuli (Kreiman et al., 2007; Maryn & Roy, 2012); however, stimuli developed for adults may not be feasible or effective for children due to task demands (Fujiki & Thibeault, 2021; Kelchner et al., 2010). Children's errors and modifications in production may limit or alter auditory-perceptual assessments and impede the usefulness of acoustic analyses. Thus, a lack of standardized stimuli to elicit voice samples adapted for children has constituted a

serious barrier to performing reliable auditory-perceptual voice assessments in this population.

The current study developed and tested pediatric-specific stimuli for the PAPE-V across childhood. Standardized sentences were tested for accuracy of production and cepstral/spectral properties, and prompts to obtain connected speech samples were examined for the duration of phonation elicited. Children produced pediatric sentences significantly more accurately than sentences designed for adults. Cepstral/spectral measures calculated on these sentences separated children with and without voice disorders. Based on these results, a final set of PAPE-V sentences and spontaneous speech prompts were selected and tested using auditory-perceptual voice evaluations of children with and without voice disorders. Reliability of auditory-perceptual voice assessments were within similar range to that previously reported in adults, and ratings significantly separated children with and without voice disorders.

Standardized Sentences

Although multiple investigations have shortened the adult CAPE-V sentences for use with children (Fujiki & Thibeault, 2021; Kelchner et al., 2010), the ability of children to produce these sentences accurately has not been

Table 7. Means and standard deviations for raw auditory-perceptual voice ratings across children with and without voice disorders.

Parameters of voice quality	Total		Typical voices		Voice disorders		<i>p</i> value
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Overall	20.7	21.1	3.4	7.6	33.1	18.8	< .001
Breathiness	17.4	20.2	2.8	6.3	27.7	20.3	< .002
Roughness	19.5	19.5	4.4	7.7	30.3	18.2	< .003
Strain	18.8	20.8	4.1	7.3	29.2	20.9	< .004
Pitch	5.3	12.8	1.6	5.1	7.9	15.6	< .005

Table 8. Reliability indices for auditory-perceptual ratings made using pediatric stimuli, as well as the adult Consensus Auditory-Perceptual Evaluation of Voice sentences.

Pediatric stimuli					Adult sentences				
Intraclass correlation coefficients (ICCs)									
Age range, years	Parameter	ICC	95% CI		Age range, years	Parameter	ICC	95% CI	
3–17	Overall voice quality	.70	.57	.78	7–17	Overall voice quality	.59	.52	.70
	Breathiness	.67	.52	.76		Breathiness	.58	.51	.71
	Roughness	.53	.39	.67		Roughness	.49	.41	.58
	Strain	.42	.29	.57		Strain	.41	.34	.51
	Pitch	.38	.09	.39		Pitch	.32	.25	.40
	Loudness	.35	.01	.42		Loudness	.32	.26	.45
Variability scores									
Age range, years	Parameter	<i>M</i>	<i>SD</i>		Age range, years	Parameter	<i>M</i>	<i>SD</i>	<i>p</i> value
3–17	Overall voice quality	259.1	318.5		7–17	Overall voice quality	321.5	415.1	.03
	Breathiness	252.0	415.6			Breathiness	301.1	407.5	.04
	Roughness	339.6	215.1			Roughness	415.1	884.4	.16
	Strain	370.6	251.4			Strain	465.5	654.4	.03

Note. 95% CI = 95% confidence interval.

Table 9. Means and standard deviations for variability scores across voice status and sex.

Factors		Interrater variability scores			Intrarater variability scores		
		M	SD	p value	M	SD	p value
Overall voice quality							
All voices		259.1	618.5		189.1	201.5	
Voice status	Typical voice	320.6	650.9	.002*	185.2	214.1	.45
	Voice disorder	267.8	467.6		205.4	140.5	
Sex	Female	231.8	902.4	.025	178.5	180.8	.64
	Male	318.8	612.5		151.1	194.2	
Breathiness							
All voices		252.0	815.6		254.1	175.2	
Voice status	Typical voice	163.5	669.1	< .001*	150.2	168.5	.43
	Voice disorder	531.5	857.9		298.1	201.1	
Sex	Female	175.1	700.3	.138	241.6	195.9	.62
	Male	435.1	895.3		260.4	187.4	
Roughness							
All voices		339.6	215.1		268.5	200.4	
Voice status	Typical voice	293.3	207.2	< .001*	178.5	163.2	.32
	Voice disorder	408.2	217.9		274.4	174.2	
Sex	Female	289.8	192.2	.192	256.3	158.2	.52
	Male	362.9	160.1		284.1	189.3	
Strain							
All voices		370.6	251.4		265.5	176.3	
Voice status	Typical voice	359.0	265.5	.089	256.3	163.1	.48
	Voice disorder	387.7	230.1		268.7	196.2	
Sex	Female	333.5	213.4	.831	250.4	198.3	.62
	Male	388.0	268.0		263.9	164.2	

*Indicates statistical significance.

established. Findings from this study confirm the hypothesis that the CAPE-V sentences are difficult for young children to produce. For children ≤ 6 years of age, accuracy on the adult sentences never reached 40% and dipped as low as 3% for the sentence “Peter will keep at the peak.” This is likely the result of the unclear morphological content of this sentence. This particular sentence has been replaced in the revised version of the CAPE-V (Kempster et al., 2025); however, it should be noted that accuracy remained low even across the more morphologically straightforward sentences. For children 7–10 years of age, accuracy generally hovered around 60% on the adult sentences, meaning nearly half of children required multiple tries to produce them. It was a goal of this study to devise sentences that children could produce on an initial try, and this was the case over 80% of the time. If children made errors or modifications, they corrected them on a second try in nearly all cases. It should be noted, however, that teenagers produced both the adult and pediatric sentences quite accurately, suggesting that either the adult or the child stimuli could be effective with adolescents. Using pediatric stimuli across childhood would promote continuity of care.

Cepstral/spectral measures calculated on the PAPE-V sentences were consistent with previous findings. CPP values for normophonic children fell within the typical range suggested by past investigations (Infusino et al., 2015; Joshi et al., 2025; Kent et al., 2021). Additionally, CPP values for children with dysphonia were significantly smaller than those with typical voices—reflecting poorer voice quality (Aghajanzadeh & Saeedi, 2022; Aydinli et al., 2019). The relationships observed between sentence types matched those reported in adults—namely, that the soft glottal attack and nasal consonant sentences displayed smaller CPP values than the other sentences (Watts, 2015). Current CPP findings also support past studies regarding the effect of age and sex on these measures in children (Demirci et al., 2021; Infusino et al., 2015; Kent et al., 2021). Specifically, CPP increases with age (Heller Murray & Yucel, 2024), and sex differences emerge in adolescents—with male adolescents displaying significantly larger values than female adolescents (Fujiki et al., 2023; Joshi et al., 2025).

Current LHR findings are generally consistent with the past study (Fujiki, Venkatraman, & Heller Murray, 2025). It has been previously documented that LHR values increase across childhood; however, these changes are of smaller magnitude and more gradual than those observed with CPP (Demirci et al., 2021; Infusino et al., 2015; Kent et al., 2021). Research suggests that LHR may not be as sensitive as CPP to the presence of benign vocal fold lesions in children (Fujiki et al., 2023). In the current study, benign lesions tempered the typical increases in CPP associated with development. This is likely because benign lesions inhibit glottic closure and increase vocal breathiness

(Gramuglia et al., 2014), thereby masking the subtle decreases in breathiness, which typically occur as children age.

It should be noted that no significant differences in cepstral/spectral measures were observed across the PAPE-V sentences and the adult CAPE-V sentences. This would indicate that either stimulus set should produce similar acoustic results. It should be remembered that children ≤ 6 years of age did not produce the adult stimuli with sufficient accuracy to be included in this analysis; thus, the pediatric stimuli would be preferable in young children.

Spontaneous Speech Prompts

The ideal speech prompt should efficiently produce as much phonation as possible. As such, longer duration samples with maximum syllables produced are ideal. For adolescents, pediatric and adult spontaneous speech prompts elicited similar voice samples using all tested questions. For younger children, however, samples varied across prompts. Overall, the prompts, “Tell me about your family” and “Tell me how to make a peanut butter and jelly sandwich,” were most effective in eliciting connected speech. In contrast, asking children to describe the Cookie Theft picture was less efficient as children often paused to scan the picture when responding. Similarly, asking “What is your favorite movie?” often elicited minimal responses if children did not watch movies, so this prompt was not included in the final version. No single prompt was 100% effective across all children, but most children produced sufficient samples without follow-up questions. It is recognized, however, that all prompts, especially “Tell me how to make a peanut butter and jelly sandwich,” may need to be adjusted to reflect a child’s background, culture, and experience. A clinician may ask a child to describe how any favorite food is made if they are not familiar with peanut butter and jelly sandwiches.

Children’s behavior in medical assessments may be variable and often difficult to predict. Factors such as shyness, fatigue, distraction, or negative emotion may complicate speech sample elicitation no matter how effective a particular prompt might be. As such, there is likely no spontaneous speech prompt that will work in 100% of situations, particularly with young children. In the current study, however, the PAPE-V sentences and prompts elicited speech samples that were sufficient for auditory-perceptual voice assessment for children in all of the age groups.

Auditory-Perceptual Assessment Reliability

Raw auditory-perceptual ratings were significantly higher for children with voice disorders across all parameters. This indicates that overall, clinicians perceived that children with benign vocal fold lesions sounded more dysphonic than vocally healthy children. This was expected given that

laryngeal pathology often results in dysphonic voice quality (Aoki et al., 2021; Gramuglia et al., 2014). As clinicians were blinded to each child's vocal status, it was encouraging that ratings reflected the worsening of voice quality that was also captured by cepstral/spectral measures.

Overall, ICCs were in similar range to those previously reported in adults using the CAPE-V in comparable numbers of voices and raters. Across all voices, study findings confirmed past work suggesting that overall voice quality is most reliably rated, and ratings of vocal strain are most variable (Dahl et al., 2021; Zraick et al., 2011). In the current study, ICCs were lower than those previously reported for adults across most parameters of voice with the exception of breathiness, which was actually higher than reported in past work (Zraick et al., 2011). Overall, considering that relatively little is understood about what constitutes "typical" voice quality in children, it is encouraging that rating reliability was relatively high for children across the wide range of ages examined in this study. It remains clear, however, that future study is warranted.

Variability analyses confirmed that overall voice quality and breathiness were significantly more reliably rated than roughness or strain. This might be expected since it can be difficult to define specific vocal parameters (Eadie & Doyle, 2002; Kreiman & Gerratt, 2000), and past studies in adults indicate that overall voice quality is more reliably rated than individual parameters of voice (Dahl et al., 2021; Zraick et al., 2011). Interestingly, in children, variability scores for breathiness were in a similar range to overall voice quality—suggesting that this vocal parameter was rated with comparable reliability. This may be because breathiness tends to be the most common vocal deviation in children (Oliveira et al., 2011). Additionally, breathiness may be easier to conceptualize than other parameters of voice quality as this precept not only corresponds with acoustic measures but can also have direct physiological underpinnings (Park et al., 2019). In fact, auditory-perceptual ratings of breathiness strongly correlate with acoustic voice measures in children with dysphonia (Fujiki & Thibeault, 2021). This suggests that breathiness may be one of the most salient features of pediatric voice disorders.

Ratings of overall voice quality were significantly more reliable for children with voice disorders. This may reflect the fact that auditory-perceptual voice assessments have rarely been studied in children with typical voices, and it is unclear what level of breathiness, roughness, or strain might be expected. In fact, preliminary evidence indicates that among normophonic school-aged children, varying levels of breathiness and roughness may be perceived by listeners (Aires et al., 2019). Thus, rater conceptualization of a typical pediatric voice might have been somewhat individual (Kreiman et al., 2021). It is also

possible that measuring weak percepts is difficult (Bele, 2005). Thus, small amounts of breathiness or strain in normophonic voices may be more difficult to reliably quantify. Still, overall voice quality was rated fairly reliably across children with and without voice disorders. That is, SLPs generally agreed about what sounded typical—and this usually aligned with voice status.

For breathiness and roughness, ratings were more reliable for normophonic children when compared to those with benign lesions. This likely reflects the fact that although there was variation in what SLPs considered to be normal voice quality, children with typical voices were generally rated fairly low on these parameters. In contrast, for children with voice disorders, ratings of breathiness and roughness were more varied and of larger magnitude. Thus, interrater reliability may have been lower because of greater variation in children with voice disorders. It is also possible that SLP's internal standards to rate breathiness and roughness were somewhat individual (Eadie & Baylor, 2006; Gama et al., 2024; Silva et al., 2012; Yiu & Ng, 2004). Additionally, parameters such as breathiness and roughness may interact and complicate auditory-perceptual assessment (Anand, 2025; Eddins et al., 2021; Park et al., 2022). Preliminary evidence suggests that breathiness and roughness may be more reliably related when anchor voices serve as external references (dos Santos et al., 2019). This may be clinically challenging given that voices change across childhood and sex, but further investigation is warranted.

Vocal strain was the only vocal parameter to be rated with similar reliability across children with and without dysphonia. This likely reflects the fact that vocal strain has historically been more difficult to rate than other vocal parameters in adults (Solomon et al., 2011; Zraick et al., 2011). In addition, it is unclear what level of vocal strain is typical in children. Past investigation indicates that teachers may perceive similar levels of vocal strain across children with and without dysphonia (Simões-Zenari et al., 2012). Thus vocal strain is frequently present in children with vocal fold nodules (Hron et al., 2019; Nuss et al., 2010), and some level of strain may be present in typical voices as well (Fujiki & Thibeault, 2024b).

Intrarater reliability was better than interrater reliability in most cases. This would indicate that experienced SLPs had consistent internal metrics of voice quality, breathiness, roughness, and strain. No significant differences were observed between intrarater reliability in children with and without voice disorders or sex—suggesting that these internal metrics were fairly robust across populations. Future study may explore additional factors influencing intrarater reliability in larger numbers of children and raters.

Both ICC and variability scores indicated that reliability was better for auditory-perceptual voice ratings made using pediatric stimuli as opposed to the adult CAPE-V sentences; however, differences in variability scores were not statistically significant given the strict alpha utilized in this study. At minimum, this finding suggests that pediatric stimuli were equally effective to the adult sentences included in the CAPE-V. It should be remembered, however, that the youngest age group could not produce the adult sentences and auditory-perceptual assessments were therefore only performed using the pediatric stimuli. It is notable that reliability was better or similar for auditory-perceptual assessments using the pediatric sentences, and children as young as 3 years could produce these sentences. It can be concluded that the pediatric stimuli elicited an effective sample for auditory-perceptual evaluations of voice. Future work may determine whether all of the included tasks are essential (sentences, sustained vowel, spontaneous speech) and determine the minimum sample of spontaneous speech needed for reliable auditory-perceptual assessment.

Some factors should be taken into consideration when interpreting the current data. First, accuracy scores were somewhat stringent and only considered for a child's first production attempt. This was done because clinic time is frequently limited. It should be considered, however, that accuracy would be higher if analyses considered children's second attempts at sentence production. For acoustic analyses, the number of children with voice disorders was relatively small across age groups. For auditory-perceptual voice ratings, all raters were SLPs specialized in the treatment of pediatric voice. Thus reliability might well be lower among less experienced raters (Alves et al., 2024; Eadie & Kapsner-Smith, 2011; Helou et al., 2010). Future study should continue to probe the minimum duration of connected speech needed for reliable auditory-perceptual voice assessments as well as the scale type, which is most useful clinically (Anand et al., 2025).

Conclusions

PAPE-V stimuli were developed and tested for auditory-perceptual voice assessments across childhood. Pediatric sentences were produced significantly more accurately than sentences developed for adults. Accuracy was over 80% for children 3 through 17 years of age. CPP calculated on the PAPE-V sentences effectively separated children with and without voice disorders and was sensitive to the influence of age and sex. Spontaneous speech prompts were also tested and found to elicit an average of around 40 syllables of phonation over approximately 25–30 s in children 3–6 years of age. Speech samples were

longer for older children. Auditory-perceptual ratings of voice significantly separated children with and without voice disorders and displayed reliability levels similar to those previously reported for adults. For overall voice quality, reliability was significantly higher for children with benign vocal fold lesions when compared to those with typical voices. Ratings of breathiness and roughness were significantly more reliable in children with typical voices when compared to those with dysphonia. Reliability for ratings of vocal strain were similar across both groups. All findings suggest that the PAPE-V stimuli are practical and effective for auditory-perceptual assessments across childhood.

Data Availability Statement

Due to privacy and ethical concerns, data cannot be made available.

Acknowledgments

This work was funded by an ASH Foundation Clinical Research Grant (awarded to Robert Brinton Fujiki), National Institutes of Health (NIH) Grant T32-DC009401 (awarded to Susan L. Thibeault, University of Wisconsin–Madison), and NIH Grant R01-DC004336 (awarded to Susan L. Thibeault). The authors thank Maia Braden for her role in sentence development. They also acknowledge Elizabeth Heller Murray, Roseanne Clark, Loraine Syndey Kagan, Julia Leclerc, John Fredeking, Rita Patel, Christine Bueckers, Maia Braden, and Brian Peterson for their expertise in performing voice ratings. The authors thank John H. Munday, Abby Nimtz, Kate Gold Reilly, Cora Ling, and Marie de Stadler for their roles in data collection and analysis.

References

- Aghajanzadeh, M., & Saeedi, S. (2022). Efficacy of cepstral measures in voice disorder diagnosis: A literature review. *Journal of Modern Rehabilitation*, 16(2), 120–129. <https://doi.org/10.18502/jmr.v16i2.9298>
- Aires, F., Fernandes, K., Hoffmann, C. F., Finger, L. S., & Cielo, C. A. (2019). Auditory-perceptual characteristics of the voice of nondysphonic school children from 8:0 to 10:0 years old. *Journal of Voice*, 33(6), 946.e1–946.e5. <https://doi.org/10.1016/j.jvoice.2018.06.002>
- Akif Kiliç, M., Okur, E., Yildirim, I., & Güzelsoy, S. (2004). The prevalence of vocal fold nodules in school age children. *International Journal of Pediatric Otorhinolaryngology*, 68(4), 409–412. <https://doi.org/10.1016/j.ijporl.2003.11.005>
- Alves, J. do N., de Almeida, A. A. F., Yamasaki, R., & Lopes, L. W. (2024). The influence of listener experience, measurement scale and speech task on the reliability of auditory-perceptual evaluation of vocal quality. *CoDAS*, 36, Article e20230175. <https://doi.org/10.1590/2317-1782/20232023175>

- Anand, S. (2025). Perceptual and computational estimates of vocal breathiness and roughness in sustained phonation and connected speech. *Journal of Voice*, 39(4), 1131.e31–1131.e43. <https://doi.org/10.1016/j.jvoice.2023.02.014>
- Anand, S., Awan, S. N., Park, Y., Brehm, S. B., Kelchner, L., Weinrich, B., McElfresh, K., Shrivastav, R., de Alarcon, A., & Eddins, D. A. (2025). Voice quality in pediatric dysphonia: Assessing expert rater reliability across three perceptual methods. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.07.033>
- Aoki, T., Takeyama, T., Sakamoto, Y., Shimada, A., Kondo, E., Nakano, S., Fukuda, J., Azuma, T., Sato, G., Kitamura, Y., Hosokawa, K., Uda, J., & Takeda, N. (2021). Effects of vocal nodules on acoustic characteristics of voice in children: An acoustic analysis of voice. *The Journal of Medical Investigation*, 68(3.4), 276–279. <https://doi.org/10.2152/jmi.68.276>
- Aronson, A. E., & Bless, D. (2009). *Clinical voice disorders* (4th ed.). Thieme. <https://shop.thieme.com/Clinical-Voice-Disorders/9781638530497>
- Ataee, E., Khoramshahi, H., Naderifar, E., & Dastoorpour, M. (2022). Relation between Dysphonia Severity Index (DSI) and Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice*, 36(3), 435.e1–435.e14. <https://doi.org/10.1016/j.jvoice.2020.06.030>
- Awan, S. N., Roy, N., & Dromey, C. (2009). Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model. *Clinical Linguistics & Phonetics*, 23(11), 825–841. <https://doi.org/10.3109/02699200903242988>
- Aydinli, F. E., Özcebe, E., & İncebay, Ö. (2019). Use of cepstral analysis for differentiating dysphonic from normal voices in children. *International Journal of Pediatric Otorhinolaryngology*, 116, 107–113. <https://doi.org/10.1016/j.ijporl.2018.10.029>
- Bartsties, B., & De Bodt, M. (2015). Assessment of voice quality: Current state-of-the-art. *Auris Nasus Larynx*, 42(3), 183–188. <https://doi.org/10.1016/j.anl.2014.11.001>
- Bartsties, B., & Maryn, Y. (2017). The influence of voice sample length in the auditory-perceptual judgment of overall voice quality. *Journal of Voice*, 31(2), 202–210. <https://doi.org/10.1016/j.jvoice.2016.07.006>
- Bele, I. V. (2005). Reliability in perceptual analysis of voice quality. *Journal of Voice*, 19(4), 555–573. <https://doi.org/10.1016/j.jvoice.2004.08.008>
- Braden, M., & Thibeault, S. L. (2020). Outcomes of voice therapy in children with benign vocal fold lesions. *International Journal of Pediatric Otorhinolaryngology*, 136, Article 110121. <https://doi.org/10.1016/j.ijporl.2020.110121>
- Brehm, S. B., Weinrich, B., Zieser, M., Kelchner, L., Middendorf, J., Elluru, R., & de Alarcon, A. (2009). Aerodynamic and acoustic assessment in children following airway reconstruction: An assessment of feasibility. *International Journal of Pediatric Otorhinolaryngology*, 73(7), 1019–1023. <https://doi.org/10.1016/j.ijporl.2009.04.001>
- Brockmann, M., Storck, C., Carding, P. N., & Drinnan, M. J. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language, and Hearing Research*, 51(5), 1152–1160. [https://doi.org/10.1044/1092-4388\(2008\)06-0208](https://doi.org/10.1044/1092-4388(2008)06-0208)
- Calaf, N., & Garcia-Quintana, D. (2024). Development and validation of the bilingual Catalan/Spanish cross-cultural adaptation of the Consensus Auditory-Perceptual Evaluation of Voice. *Journal of Speech, Language, and Hearing Research*, 67(4), 1072–1089. https://doi.org/10.1044/2024_JSLHR-23-00536
- Campano, M., Cox, S. R., Caniano, L., & Koenig, L. L. (2023). A review of voice disorders in school-aged children. *Journal of Voice*, 37(2), 301.e1–301.e7. <https://doi.org/10.1016/j.jvoice.2020.12.018>
- Carding, P. N., Roulstone, S., Northstone, K., & ALSPAC Study Team. (2006). The prevalence of childhood dysphonia: A cross-sectional study. *Journal of Voice*, 20(4), 623–630. <https://doi.org/10.1016/j.jvoice.2005.07.004>
- Chan, K. M. K., & Yiu, E. M.-L. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research*, 45(1), 111–126. [https://doi.org/10.1044/1092-4388\(2002\)009](https://doi.org/10.1044/1092-4388(2002)009)
- Connor, N. P., Cohen, S. B., Theis, S. M., Thibeault, S. L., Heatley, D. G., & Bless, D. M. (2008). Attitudes of children with dysphonia. *Journal of Voice*, 22(2), 197–209. <https://doi.org/10.1016/j.jvoice.2006.09.005>
- Costa, C. C., Leite, A. P. D., Madazio, G., & Behlau, M. (2023). Auditory-perceptual evaluation of voice: Comparing different speech tasks to identify children with and without laryngeal lesions. *CoDAS*, 35(2), Article e20210198. <https://doi.org/10.1590/2317-1782/20212021198en>
- Dahl, K. L., Weerathunge, H. R., Buckley, D. P., Dolling, A. S., Díaz-Cádiz, M., Tracy, L. F., & Stepp, C. E. (2021). Reliability and accuracy of expert auditory-perceptual evaluation of voice via telepractice platforms. *American Journal of Speech-Language Pathology*, 30(6), 2446–2455. https://doi.org/10.1044/2021_AJSLP-21-00091
- Deborah, R., & Samayan, K. (2026). Cepstral analysis of voice in school-aged children. *Journal of Voice*, 40(4), 1245.e19–1245.e27. <https://doi.org/10.1016/j.jvoice.2024.03.019>
- Demirci, A. N., Köse, A., Aydinli, F. E., İncebay, Ö., & Yilmaz, T. (2021). Investigating the cepstral acoustic characteristics of voice in healthy children. *International Journal of Pediatric Otorhinolaryngology*, 148, Article 110815. <https://doi.org/10.1016/j.ijporl.2021.110815>
- dos Santos, P. C. M., Vieira, M. N., Sansão, J. P. H., & Gama, A. C. C. (2019). Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. *Journal of Voice*, 33(2), 220–225. <https://doi.org/10.1016/j.jvoice.2017.10.020>
- Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice*, 20(4), 527–544. <https://doi.org/10.1016/j.jvoice.2005.08.007>
- Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *The Journal of the Acoustical Society of America*, 112(6), 3014–3021. <https://doi.org/10.1121/1.1518983>
- Eadie, T. L., & Kapsner-Smith, M. (2011). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, 54(2), 430–447. [https://doi.org/10.1044/1092-4388\(2010\)09-0205](https://doi.org/10.1044/1092-4388(2010)09-0205)
- Eddins, D. A., Anand, S., Lang, A., & Shrivastav, R. (2021). Developing clinically relevant scales of breathy and rough voice quality. *Journal of Voice*, 35(4), 663.e9–663.e16. <https://doi.org/10.1016/j.jvoice.2019.12.021>
- Englert, M., Lima, L., Latoszek, B. B. v., & Behlau, M. (2022). Influence of the voice sample length in perceptual and acoustic voice quality analysis. *Journal of Voice*, 36(4), 582.e23–582.e32. <https://doi.org/10.1016/j.jvoice.2020.07.010>
- Feinstein, H., & Abbott, K. V. (2021). Behavioral treatment for benign vocal fold lesions in children: A systematic review. *American Journal of Speech-Language Pathology*, 30(2), 772–788. https://doi.org/10.1044/2020_AJSLP-20-00304
- Feinstein, H., Daşdoğan, Ü., Awan, J. A., Awan, S. N., & Abbott, K. V. (2025). Comparative analysis of two methods of perceptual voice assessment. *Journal of Voice*, 39(4), 1065–1075. <https://doi.org/10.1016/j.jvoice.2023.01.005>

- Fraila, R., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., & Gutiérrez-Arriola, J. M. (2013). Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech. *Journal of Voice*, 27(1), 11–23. <https://doi.org/10.1016/j.jvoice.2012.07.004>
- Fujiki, R. B., Braden, M., & Thibeault, S. L. (2023). Voice therapy improves acoustic and auditory-perceptual outcomes in children. *The Laryngoscope*, 133(4), 977–983. <https://doi.org/10.1002/lary.30263>
- Fujiki, R. B., & Thibeault, S. L. (2021). The relationship between auditory-perceptual rating scales and objective voice measures in children with voice disorders. *American Journal of Speech-Language Pathology*, 30(1), 228–238. https://doi.org/10.1044/2020_AJSLP-20-00188
- Fujiki, R. B., & Thibeault, S. L. (2022). Pediatric voice therapy: How many sessions to discharge? *American Journal of Speech-Language Pathology*, 31(6), 2663–2674. https://doi.org/10.1044/2022_AJSLP-22-00111
- Fujiki, R. B., & Thibeault, S. L. (2023). Examining relationships between GRBAS ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders. *Journal of Voice*, 37(3), 390–397. <https://doi.org/10.1016/j.jvoice.2021.02.007>
- Fujiki, R. B., & Thibeault, S. L. (2024a). Voice disorder prevalence and vocal health characteristics in adolescents. *JAMA Otolaryngology—Head & Neck Surgery*, 150(9), 800–810. <https://doi.org/10.1001/jamaoto.2024.2081>
- Fujiki, R. B., & Thibeault, S. L. (2024b). Voice disorder prevalence and vocal health characteristics in children. *JAMA Otolaryngology—Head & Neck Surgery*, 150(8), 677–687. <https://doi.org/10.1001/jamaoto.2024.1516>
- Fujiki, R. B., & Thibeault, S. L. (2025a). Medical comorbidities and behavioral health in adolescents with voice disorders. *JAMA Otolaryngology—Head & Neck Surgery*, 151(10), 913–922. <https://doi.org/10.1001/jamaoto.2025.2338>
- Fujiki, R. B., & Thibeault, S. L. (2025b). Socioeconomic status, voice disorder risk, and voice-related handicap across childhood. *JAMA Otolaryngology—Head & Neck Surgery*, 151(10), 967–975. <https://doi.org/10.1001/jamaoto.2025.2633>
- Fujiki, R. B., Venkatraman, A., Godbout, R. A., & Thibeault, S. L. (2025). Instrumental voice evaluation in children: What are we getting? *American Journal of Speech-Language Pathology*, 34(6), 3076–3090. https://doi.org/10.1044/2025_AJSLP-25-00161
- Fujiki, R. B., Venkatraman, A., & Heller Murray, E. S. (2025). The pediatric vocal mechanism: Structure and function. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.03.025>
- Gama, A. C. C., Mata, S. M. da, Santos, P. C. M. dos, Vieira, M. N., Sansão, J. P. H., & Quinino, R. da C. (2024). Auditory training with synthesized voice anchors: Effects on rater agreement. *Journal of Voice*, 38(2), 366–375. <https://doi.org/10.1016/j.jvoice.2021.09.009>
- Gilman, M., Shelly, S., & Gillespie, A. I. (2023). Vocal acoustics and aerodynamics during scripted reading compared to spontaneous speech. *Journal of Voice*, 37(4), 539–545. <https://doi.org/10.1016/j.jvoice.2021.03.022>
- Gómez-García, J. A., Moro-Velázquez, L., & Godino-Llorente, J. I. (2019). On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51, 181–199. <https://doi.org/10.1016/j.bspc.2018.12.024>
- Gramming, P., Sundberg, J., Ternstrom, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice*, 2(2), 118–126. [https://doi.org/10.1016/S0892-1997\(88\)80067-5](https://doi.org/10.1016/S0892-1997(88)80067-5)
- Gramuglia, A. C. J., Tavares, E. L. M., Rodrigues, S. A., & Martins, R. H. G. (2014). Perceptual and acoustic parameters of vocal nodules in children. *International Journal of Pediatric Otorhinolaryngology*, 78(2), 312–316. <https://doi.org/10.1016/j.ijporl.2013.11.032>
- Hartley, N. A., Braden, M., & Thibeault, S. L. (2017). Practice patterns of speech-language pathologists in pediatric vocal health. *American Journal of Speech-Language Pathology*, 26(2), 281–300. https://doi.org/10.1044/2016_AJSLP-15-0057
- Heller Murray, E., & Yucel, R. (2024). Longitudinal evaluation of cepstral peak prominence in children. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2024.04.019>
- Helou, L. B., Solomon, N. P., Henry, L. R., Coppit, G. L., Howard, R. S., & Stojadinovic, A. (2010). The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. *American Journal of Speech-Language Pathology*, 19(3), 248–258. [https://doi.org/10.1044/1058-0360\(2010\)09-0012](https://doi.org/10.1044/1058-0360(2010)09-0012)
- Heman-Ackah, Y. D., Heuer, R., Michael, D., Ostrowski, R., Horman, M., Baroody, M., Hilenbrand, J., & Sataloff, R. (2003). Cepstral peak prominence: A more reliable measure of dysphonia. *Annals of Otolaryngology, Rhinology & Laryngology*, 112(4), 324–333. <https://doi.org/10.1177/000348940311200406>
- Hillman, R. (2003). *Evaluation of Voice (CAPE-V) Instrument Developed by ASHA Special Interest Division 3* [Paper presentation]. The Voice Foundation's 32nd Annual Symposium: Care of the Professional Voice.
- Hron, T. A., Kavanagh, K. R., & Murray, N. (2019). Diagnosis and treatment of benign pediatric lesions. *Otolaryngologic Clinics of North America*, 52(4), 657–668. <https://doi.org/10.1016/j.otc.2019.03.010>
- Ilić-Savić, I., & Petrović-Lazić, M. (2023). The influence of voice and speech disorders on the quality of life of school-age children. *Research in Pedagogy*, 13(1), 88–97. https://research.rs/wp-content/uploads/2023/06/2217-7337_v13_n01_p088.pdf [PDF]
- Infusino, S. A., Diercks, G. R., Rogers, D. J., Garcia, J., Ojha, S., Maurer, R., Bunting, G., & Hartnick, C. J. (2015). Establishment of a normative cepstral pediatric acoustic database. *JAMA Otolaryngology—Head & Neck Surgery*, 141(4), 358–363. <https://doi.org/10.1001/jamaoto.2014.3545>
- Joshi, A., Awan, S. N., Rubino, M., Devore, D., Procter, T., & Ongkasuwan, J. (2025). Cepstral peak prominence in nondysphonic children using Praat and analysis of dysphonia in speech and voice. *Journal of Speech, Language, and Hearing Research*, 68(8), 3733–3747. https://doi.org/10.1044/2025_JSLHR-25-00046
- Kelchner, L. N., Brehm, S. B., de Alarcon, A., & Weinrich, B. (2012). Update on pediatric voice and airway disorders: Assessment and care. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 20(3), 160–164. <https://doi.org/10.1097/MOO.0b013e3283530ecb>
- Kelchner, L. N., Brehm, S., Weinrich, B., Middendorf, J., de Alarcon, A., Levin, L., & Elluru, R. (2010). Perceptual evaluation of severe pediatric voice disorders: Rater reliability using the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice*, 24(4), 441–449. <https://doi.org/10.1016/j.jvoice.2008.09.004>
- 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(2), 124–132. [https://doi.org/10.1044/1058-0360\(2008\)08-0017](https://doi.org/10.1044/1058-0360(2008)08-0017)

- Kempster, G. B., Nagle, K. F., & Solomon, N. P. (2025). Development and rationale for the Consensus Auditory-Perceptual Evaluation of Voice-Revised (CAPE-Vr). *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.01.022>
- Kent, R. D., Eichhorn, J. T., & Vorperian, H. K. (2021). Acoustic measures of voice in typically developing children ages 4 to 19 years. *International Journal of Pediatric Otorhinolaryngology*, 142, Article 110614. <https://doi.org/10.1016/j.ijporl.2021.110614>
- Kreiman, J., Auszmann, A., & Gerratt, B. R. (2021). What does it mean for a voice to sound “normal”? In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers* (pp. 83–99). Springer. https://doi.org/10.1007/978-981-15-6627-1_5
- Kreiman, J., & Gerratt, B. R. (2000). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America*, 108(4), 1867–1876. <https://doi.org/10.1121/1.1289362>
- Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*, 122(4), 2354–2364. <https://doi.org/10.1121/1.2770547>
- Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, 36(1), 21–40. <https://doi.org/10.1044/jshr.3601.21>
- Lass, N. J., Ruscello, D. M., Bradshaw, K. H., & Blankenship, B. L. (1991). Adolescents’ perceptions of normal and voice-disordered children. *Journal of Communication Disorders*, 24(4), 267–274. [https://doi.org/10.1016/0021-9924\(91\)90002-Z](https://doi.org/10.1016/0021-9924(91)90002-Z)
- Lass, N. J., Ruscello, D. M., Stout, L. L., & Hoffman, F. M. (1991). Peer perceptions of normal and voice-disordered children. *Folia Phoniatrica et Logopaedica*, 43(1), 29–35. <https://doi.org/10.1159/000266098>
- Law, T., Kim, J. H., Lee, K. Y., Tang, E. C., Lam, J. H., van Hasselt, A. C., & Tong, M. C. (2012). Comparison of rater’s reliability on perceptual evaluation of different types of voice sample. *Journal of Voice*, 26(5), 666.e13–666.e21. <https://doi.org/10.1016/j.jvoice.2011.08.003>
- Lechian, J. R., Morsomme, D., Finck, C., Huet, K., Delvaux, V., Piccaluga, M., Harmegnies, B., & Saussez, S. (2018). The effect of the speech task characteristics on perceptual judgment of mild to moderate dysphonia: A methodological study. *Folia Phoniatrica et Logopaedica*, 70(3–4), 156–164. <https://doi.org/10.1159/000492219>
- Ma, E. P.-M., & Yu, C. H.-Y. (2013). Listeners’ attitudes toward children with voice problems. *Journal of Speech, Language, and Hearing Research*, 56(5), 1409–1415. [https://doi.org/10.1044/1092-4388\(2013/11-0242\)](https://doi.org/10.1044/1092-4388(2013/11-0242))
- Mahalingam, S., Venkatraman, Y., & Boominathan, P. (2024). Cross-cultural adaptation and validation of Consensus Auditory Perceptual Evaluation of Voice (CAPE-V): A systematic review. *Journal of Voice*, 38(3), 630–640. <https://doi.org/10.1016/j.jvoice.2021.10.022>
- Maryn, Y., & Roy, N. (2012). Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *Journal Da Sociedade Brasileira de Fonoaudiologia*, 24(2), 107–112. <https://doi.org/10.1590/S2179-64912012000200003>
- Masaki, A. (2001). *Optimizing acoustic and perceptual assessment of voice quality in children with vocal nodules* [Doctoral Dissertation, Massachusetts Institute of Technology]. MIT Open Scholarship. <https://dspace.mit.edu/entities/publication/beb5473a-5e03-4866-b3e0-bd4640de3e2c>
- Maturo, S., Hill, C., Bunting, G., Ballif, C., Maurer, R., & Hartnick, C. (2012). Establishment of a normative pediatric acoustic database. *JAMA Otolaryngology—Head & Neck Surgery*, 138(10), 956–961. <https://doi.org/10.1001/2013.jamaoto.104>
- Murton, O., Hillman, R., & Mehta, D. (2020). Cepstral peak prominence values for clinical voice evaluation. *American Journal of Speech-Language Pathology*, 29(3), 1596–1607. https://doi.org/10.1044/2020_AJSLP-20-00001
- Nardone, H. C., Recko, T., Huang, L., & Nuss, R. C. (2014). A retrospective review of the progression of pediatric vocal fold nodules. *JAMA Otolaryngology—Head & Neck Surgery*, 140(3), 233–236. <https://doi.org/10.1001/jamaoto.2013.6378>
- Nemr, K., Simoes-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). GRBAS and CAPE-V scales: High reliability and consensus when applied at different times. *Journal of Voice*, 26(6), 812.e17–812.e22. <https://doi.org/10.1016/j.jvoice.2012.03.005>
- Nuss, R. C., Ward, J., Huang, L., Volk, M., & Woodnorth, G. H. (2010). Correlation of vocal fold nodule size in children and perceptual assessment of voice quality. *Annals of Otolaryngology & Laryngology*, 119(10), 651–655. <https://doi.org/10.1177/000348941011901001>
- Oliveira, R. C., Teixeira, L. C., Gama, A. C. C., & Medeiros, A. M. de. (2011). Auditory-perceptive, acoustic and vocal self-perception analyses in children. *Jornal Da Sociedade Brasileira de Fonoaudiologia*, 23, 158–163. <https://doi.org/10.1590/S2179-64912011000200013>
- Park, Y., Anand, S., Kopf, L. M., Shrivastav, R., & Eddins, D. A. (2022). Interactions between breathy and rough voice qualities and their contributions to overall dysphonia severity. *Journal of Speech, Language, and Hearing Research*, 65(11), 4071–4084. https://doi.org/10.1044/2022_JSLHR-22-00012
- Park, Y., Perkell, J. S., Matthies, M. L., & Stepp, C. E. (2019). Categorization in the perception of breathy voice quality and its relation to voice production in healthy speakers. *Journal of Speech, Language, and Hearing Research*, 62(10), 3655–3666. https://doi.org/10.1044/2019_JSLHR-S-19-0048
- Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Delyski, D., Eadie, T., Paul, D., Švec, J. G., & Hillman, R. (2018). Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *American Journal of Speech-Language Pathology*, 27(3), 887–905. https://doi.org/10.1044/2018_AJSLP-17-0009
- Pommée, T., Mbagira, D., & Morsomme, D. (2026). French-language adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice*, 40(4), 1239.e21–1239.e32. <https://doi.org/10.1016/j.jvoice.2024.03.011>
- Reynolds, V., & Fleury, A. (2024). A preliminary exploration of conversation sampling in pediatric voice disorders. *Logopedics Phoniatrics Vocology*, 49(1), 41–46. <https://doi.org/10.1080/14015439.2022.2102207>
- Reynolds, V., Scavo-Smith, K., Oteng-Bediako, K., & Scanlon, S. (2025). An observational study of discourse tasks and running speech sampling in the assessment of paediatric voice quality. *International Journal of Language & Communication Disorders*, 60(6), Article e70132. <https://doi.org/10.1111/1460-6984.70132>
- Roth, C. (2011). Boston Diagnostic Aphasia Examination. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp. 428–430). Springer. https://doi.org/10.1007/978-0-387-79948-3_868
- Sandage, M. J., Plexico, L. W., & Schiwitz, A. (2015). Clinical utility of CAPE-V sentences for determination of speaking

- fundamental frequency. *Journal of Voice*, 29(4), 441–445. <https://doi.org/10.1016/j.jvoice.2014.09.005>
- Silva, R. S. A., Simões-Zenari, M., & Nemr, N. K.** (2012). Impact of auditory training for perceptual assessment of voice executed by undergraduate students in speech-language pathology. *Jornal Da Sociedade Brasileira de Fonoaudiologia*, 24(1), 19–25. <https://doi.org/10.1590/S2179-64912012000100005>
- Simões-Zenari, M., Nemr, K., & Behlau, M.** (2012). Voice disorders in children and its relationship with auditory, acoustic and vocal behavior parameters. *International Journal of Pediatric Otorhinolaryngology*, 76(6), 896–900. <https://doi.org/10.1016/j.ijporl.2012.02.067>
- Solomon, N. P., Helou, L. B., & Stojadinovic, A.** (2011). Clinical versus laboratory ratings of voice using the CAPE-V. *Journal of Voice*, 25(1), e7–e14. <https://doi.org/10.1016/j.jvoice.2009.10.007>
- Verduyck, I., Remacle, M., Jamart, J., Benderitter, C., & Morsomme, D.** (2011). Voice-related complaints in the pediatric population. *Journal of Voice*, 25(3), 373–380. <https://doi.org/10.1016/j.jvoice.2009.11.008>
- Watts, C. R.** (2015). The effect of CAPE-V sentences on cepstral/spectral acoustic measures in dysphonic speakers. *Folia Phoniatrica et Logopaedica*, 67(1), 15–20. <https://doi.org/10.1159/000371656>
- Watts, C. R., & Awan, S.** (2011). Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research*, 54(6), 1525–1537. [https://doi.org/10.1044/1092-4388\(2011/10-0209\)](https://doi.org/10.1044/1092-4388(2011/10-0209))
- Yiu, E. M.-L., & Ng, C.-Y.** (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics*, 18(3), 211–229. <https://doi.org/10.1080/0269920042000193599>
- Zraick, R. I., Kempster, G. B., Connor, N. P., Thibault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E.** (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology*, 20(1), 14–22. [https://doi.org/10.1044/1058-0360\(2010/09-0105\)](https://doi.org/10.1044/1058-0360(2010/09-0105))