

Research Article

Pediatric Auditory-Perceptual Evaluation of Voice Part 2: A Protocol

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

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

ARTICLE INFO

Article History:

Received November 12, 2025

Revision received March 30, 2026

Accepted April 6, 2026

Editor-in-Chief: Jessica E. Huber

Editor: Amanda I. Gillespie

https://doi.org/10.1044/2026_JSLHR-25-00924

ABSTRACT

Purpose: Although auditory-perceptual voice assessments are crucial for evaluating pediatric voices, no protocol designed for use with children has been established. This study created and validated the Pediatric Auditory-Perceptual Evaluation of Voice (PAPE-V) protocol for use across childhood. Influence of scale type and speaker demographics on voice rating reliability was considered.

Method: Ten speech-language pathologists (SLPs) specialized in pediatric voice rated voice samples from 102 children aged 3–17 years with and without diagnosed voice disorders (47% female). SLPs were asked to rate each voice for overall voice quality, breathiness, roughness, strain, pitch, and loudness. Voices were rated twice across two separate sessions, with 20% of voices repeated to allow for the calculation of intrarater reliability. In each session, either visual analog scales (VASs) or Likert scales were utilized (order counterbalanced across raters). Interrater and intrarater reliabilities were compared across rating scales and child demographics using intraclass correlation coefficients (ICCs) and Krippendorff's alphas. SLPs' feedback regarding both scale types was also obtained.

Results: Interrater and intrarater reliabilities were high across all conditions. For both scale types, ICCs were highest for overall voice quality (Likert ICC = .97, VAS ICC = .96) and lowest for pitch (Likert ICC = .82, VAS ICC = .78). Interrater reliability was significantly greater for children presenting with voice disorders than for children with healthy voices ($p < .01$). Interrater reliability also varied with age, and ratings for children younger than 6 years had significantly better interrater reliability than ratings for children 7–17 years old ($p < .001$). No significant differences in reliability were observed across scale type or sex. Clinicians indicated that Likert scales were more time efficient than VASs; however, either scale type could be used in the PAPE-V.

Conclusions: The PAPE-V can be used to perform reliable auditory-perceptual voice assessments in children using either VAS or Likert scales. This protocol constitutes an important step in promoting the reliability and efficiency in pediatric voice assessments.

Pediatric voice disorders constrain communication and undermine children's ability to share their experience and express their daily needs (Cohen & Wynne, 2015; Fujiki & Thibeault, 2025b). Unfortunately, voice disorders are fairly common across childhood. In fact, a recent study of the U.S. population suggests that 6.7% of school-aged

children and 7.4% of adolescents report current voice problems (Fujiki & Thibeault, 2024b, 2024a). Additionally, lifetime voice disorder prevalence in school-aged children is 12% and as high as 24.3% in adolescents aged 13–17 years (Fujiki & Thibeault, 2024b, 2024a). Voice disorders generally manifest as dysphonia or disruptions in voice quality, pitch, or loudness, which fail to meet a person's needs and/or engender negative attention given an individual's sex, age, or gender (Aronson & Bless, 2009). Best practice dictates that voice quality in children be assessed using a combination of

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.

auditory-perceptual, acoustic, aerodynamic, and self-reported voice measures (Patel, 2023). It is often the case, however, that children can only tolerate limited voice assessment procedures in the clinic (Fujiki, Venkatraman, Godbout, & Thibeault, 2025), and they may not be able to perform the tasks necessary for quantitative voice measures due to developmental, behavioral, or physiological immaturity. Even when quantitative voice measures can be collected, there is no single outcome that can capture the nuances perceived by expert listeners (Kent, 1999). Fortunately, auditory-perceptual voice assessments can be effective for evaluating dysphonia holistically and functionally (Eadie & Doyle, 2005; Kreiman et al., 1993). As such, an evidence-based auditory-perceptual voice assessment protocol designed for children is needed.

Auditory-perceptual evaluation of voice involves listening to an individual's voice in order to assess and document overall voice quality as well as specific parameters of voice such as roughness, breathiness, or strain (Walden & Rau, 2022). Additionally, pitch and loudness must be assessed in reference to a child's sex, gender, and age (Fujiki, Venkatraman, & Heller Murray, 2025). In the treatment of voice disorders, a speech-language pathologist (SLP) is usually primarily responsible for conducting auditory-perceptual voice assessments. SLPs frequently perform these evaluations to assess the severity of vocal deviation (Awan et al., 2010) and to gauge response to treatment (Akin Şenkal & Çiyiltepe, 2013; Fujiki et al., 2023; Hseu et al., 2025). It is essential, however, that auditory-perceptual voice assessments be reliable; that is, auditory-perceptual voice ratings should be performed consistently across raters and time points. The reliability of auditory-perceptual voice ratings has been the subject of considerable study; however, previous investigations have almost exclusively examined adult populations (Anand, 2025; Benoy & Jayakumar, 2026; Dahl et al., 2021; do Alves et al., 2024; Eadie & Kapsner-Smith, 2011; Fujiki et al., 2022; Sauder et al., 2024; Walden, 2022). It remains unclear, then, what factors influence the reliability of auditory-perceptual ratings for children.

No child-specific protocol for auditory-perceptual voice evaluation has been reported in the literature. As a result, these assessments have generally been completed using protocols designed for adults (Braden & Thibeault, 2020; Hseu et al., 2023; Kelchner et al., 2010). The two most commonly employed measures are the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster et al., 2009), which has a newly published revised version known as the CAPE-Vr, and the Grade Roughness Breathiness Asthenia and Strain (GRBAS) scales (Hirano, 1981). Although development of these protocols was not data driven, both have been subsequently studied in a variety of research and clinical contexts. The CAPE-V standardizes the stimuli on which voice ratings are performed. Overall voice

quality, roughness, breathiness, and strain are rated on visual analog scales (VASs) anchored by *normal* (0) on one end and *extreme* (100) on the other. The GRBAS scale consists of a 4-point Likert scale on which SLPs rate overall voice quality, in addition to roughness, breathiness, asthenia, and strain (Hirano, 1981). Possible ratings include 0 (*normal*), 1 (*mild*), 2 (*moderate*), and 3 (*severe*). The CAPE-Vr was created using clinician feedback; however, data regarding its use are not available due to how recently it was published. Voice ratings made using the CAPE-V and the GRBAS scales have been shown to display good intrarater and interrater reliabilities when assessing adult populations (Zraick et al., 2011). Additionally, voice ratings made from both protocols are correlated in children (Fujiki & Thibeault, 2021) and have been used to assess the effects of laryngeal pathology (Costa et al., 2023; Rossi et al., 2026) and voice therapy in pediatric populations (Fujiki et al., 2023; Fujiki & Thibeault, 2022; Salderay et al., 2024; Tezcaner et al., 2009). It remains unclear, however, whether these rating protocols are optimized for use with pediatric voices.

Although the CAPE-V and GRBAS have been used with children in research and clinical contexts (Gramuglia et al., 2014; Reynolds et al., 2025), preliminary investigation suggests that aspects of these protocols may be problematic for this population (Fujiki & Thibeault, 2021; Kelchner et al., 2010). For example, CAPE-V stimuli are often too difficult for children to produce due to the cognitive and linguistic demands of the tasks. As such, past studies have shortened the stimuli informally, potentially leading to differing protocols across institutions and settings (Fujiki & Thibeault, 2021; Kelchner et al., 2010). Additionally, in actual clinical situations, the CAPE-V is often not utilized as prescribed due to time constraints (Lodhavia & Kempster, 2026). As an example, some clinicians do not use the VASs as intended but rather assign voices a number from 1 to 100 based on their general impressions (Nagle et al., 2024). This practice functionally converts the VAS to a 100-point Likert scale, thus influencing voice rating psychometrics. The resulting ratings may be less reliable than those using actual VASs. It is also the case, however, that the full CAPE-V protocol may not always be feasible in clinical situations.

The GRBAS scale may be administered fairly quickly; however, there is virtually no information precisely clarifying how the scale should be employed. As such, use of the GRBAS scale likely varies across institutions and raters. As a result, voice ratings may be difficult to compare across contexts, thus disrupting continuity of care. Additionally, the limited rating options on the GRBAS scale tend to cluster pediatric patients into the “moderate” category, leaving a wide range of dysphonia under a single rating (Fujiki & Thibeault, 2021). It is unclear whether this scale

is sufficiently sensitive to the wide range of pediatric voices, which present in clinical practice.

It should also be noted that other scale types such as paired comparison, direct magnitude estimation (DME), and visual sort and rank have been used to assess pediatric voices in research contexts (Anand et al., 2025; Feinstein et al., 2025). Preliminary work suggests that VASs have benefits over paired comparison ratings, as they are more consistent and allow for more detail regarding vocal parameters (Feinstein et al., 2025). Initial study indicates that visual rate and sort and DME scales may have similar reliability to VASs (Anand et al., 2025; Eadie & Doyle, 2005; Y. Lee et al., 2025); however, clinical implementation of these scales is challenging. Anchor voices may be needed, which can be difficult to establish given the variability inherent in pediatric voices across childhood. Additionally, these methods may require a clinician to listen to multiple voices, thereby increasing the time needed to perform auditory-perceptual assessment. Thus, at this time, VAS or Likert scales are most practical for clinical contexts.

Considering that the cognitive, developmental, and vocal functioning of children differs from that of adults, an auditory-perceptual voice assessment protocol designed specifically for children was developed. Initially, standardized voice elicitation stimuli were developed and tested for use with children (Fujiki et al., 2026). Elicitation stimuli included connected speech to reflect how children phonate in daily communicative contexts (Kreiman & Gerratt, 2000; Law et al., 2012; Lechien et al., 2018; Maryn & Roy, 2012) and a sustained vowel. Stimuli validation ensured that standardized sentences could be accurately produced across childhood and that auditory-perceptual assessments made using these stimuli were similarly reliable (or more reliable) than those made using the adult CAPE-V sentences. Additionally, cepstral/spectral analyses were performed on standardized sentences to ensure that acoustic measures calculated on these stimuli separated pediatric dysphonic and normophonic voices (Fujiki et al., 2026).

In addition to developing stimuli appropriate for children, it was important to explore the scale rating type that would be most useful in rating children's responses to the standardized stimuli. Determining the scale type for child assessments can be a complex endeavor; however, the psychometrics of pediatric voice ratings have rarely been examined. Although past studies have examined voice ratings made on a variety of scale types for adults (Ataee et al., 2022; Choi et al., 2021; De Bodt et al., 1997; Fujiki & Thibeault, 2023; Karnell et al., 2007; Labaere et al., 2024; Nagle et al., 2014; Ricci-Maccarini et al., 2022; Tunç Songur et al., 2025; Wischhoff et al., 2025), the CAPE-V and GRBAS are most commonly employed. For adults, VASs such as those used in the CAPE-V have frequently been reported to

be more reliable than ratings made on the GRBAS or other equal appearing interval (EAI) scales (Baravieira et al., 2016; Kreiman et al., 2007; Martins et al., 2015; Yamasaki et al., 2017; Zraick et al., 2011). Additionally, data from VASs are more conducive to statistical analysis when compared with ordinal data produced by the GRBAS (Fujiki & Thibeault, 2021; Nemr et al., 2012), which may facilitate the comparison of voice quality across multiple time points.

VASs, however, are not without their drawbacks. Although VASs allow for more fine-tuned ratings of voice quality, the large range of options may lead to poorer reliability than that found in scales with fewer options (Wuyts et al., 1999). Additionally, this scale type may cause end-aversion, which occurs when raters are hesitant to rate samples near the extremities of the line (Walden & Rau, 2022). Perhaps most worrisome, it is often unclear what any specific value on the VAS might indicate clinically. The original CAPE-V scale included mild, moderate, and severe markers, which were intended as regions of severity as opposed to cut-off scores. These were removed from the revised version of the scale; however, as they proved more problematic than helpful for clinicians (Kempster et al., 2025). As such, the clinical significance of any given value is often unclear. Although scales such as the GRBAS may be faster to use, limited options may render these scales less sensitive to incremental improvements or minor deteriorations in voice quality. Additionally, Likert-type scales generally produce ordinal data, which can be more difficult to analyze statistically. This can be challenging for clinicians and researchers looking to identifying patterns or changes in voice quality over time.

The purpose of the current study was to develop a reliable evaluation protocol for auditory-perceptual voice assessment in children. This protocol will henceforth be referred to as the Pediatric Auditory-Perceptual Evaluation of Voice (PAPE-V) protocol. A previous investigation developed and tested auditory-perceptual voice evaluation stimuli designed specifically for the pediatric population. The current study extended this work by utilizing these evaluation stimuli and probing rating scale types in children. The PAPE-V protocol was administered to children with and without voice disorders in order to assess how the protocol functioned across populations. Since the intent of this study was that the PAPE-V would be effective across childhood, the influence of age and sex on auditory-perceptual voice rating reliability was considered.

Method

Study Design

This study created and validated the PAPE-V protocol. Stimuli were developed, tested, and reported in a previous

investigation (Fujiki et al., 2026). The current study addressed scale type, rating psychometrics, and dimensionality across age and sex in children with and without diagnosed voice disorders. Ten SLPs specialized in the diagnosis and treatment of pediatric dysphonia rated the voices of 102 children with and without voice disorders. Voices were rated twice across two separate sessions approximately 2–3 months apart. In each session, either VASs or Likert scales were utilized, with presentation order being counterbalanced across raters. Clinician feedback regarding both scale types was obtained. All procedures were approved by both the University of Wisconsin–Madison (No. 2023-0227) and the Indiana University School of Medicine (No. 25281) institutional review boards.

Participants

Children

Three- to 17-year-old children with and without voice disorders were recruited for this study. For children with voice disorders, recruitment occurred through the outpatient voice and swallow clinic of a large tertiary children's hospital. For children without voice disorders, recruitment occurred through the otolaryngology clinic of the same hospital as well as the general community. For both groups, children were required to meet the following criteria for inclusion: (a) be between the ages of 3 and 17 years, (b) speak fluent English, and (c) be diagnosed with a voice disorder or deny any history of voice problems. For inclusion in the voice disorder group, children were required to have undergone laryngeal imaging and be diagnosed with a voice disorder by a pediatric otolaryngologist. Children in the vocally healthy group did not undergo laryngeal imaging if it was not medically indicated; however, both the child and their caregiver denied any history of voice complaints. Additionally, voices were screened by the authors for gross deviations in vocal quality. Children recruited through the otolaryngology clinic were being seen for evaluations of the tonsil/adenoids, disordered sleep breathing, cerumen removal, or mild dysphagia. Participants were not included if any voice-related symptoms were reported. Three children were later determined to present with hypernasal speech, despite typical voice quality. These individuals were included because variations in resonance may be encountered clinically. Exclusion criteria for both groups included the following: (a) head or neck cancer, (b) respiratory conditions impacting phonation, (c) speech dysarthria, and (d) neurological impairments preventing assent. Children were divided into three separate age groups as follows: Group 1 = children ≤ 6 years of age, Group 2 = children 7–10 years of age, and Group 3 = children 11–17 years of age. Age groups were created based on cognitive and vocal development (Fujiki, Venkatraman, & Heller Murray, 2025). Past study suggests that vocal fold structure becomes more defined around the age of 7 years

(Hartnick et al., 2005) but that organized cell populations defining superficial, intermediate, and deep lamina propria layers are not consistently developed until around the age of 11 years (Hirano et al., 1983). As such, age cutoffs were placed at 7 and 11 years of age. Caregivers gave informed consent, and children were assented prior to participation.

SLP Raters

SLPs specialized in the treatment and diagnosis of pediatric voice disorders were recruited from across the United States. SLPs were required to be clinically certified and primarily employed in the treatment and/or research of pediatric voice. Recruitment occurred via e-mails sent to pediatric voice clinics and research labs throughout the country. All SLPs (a) had a minimum of 5 years' experience working with pediatric voice, (b) were native English speakers, (c) were professionally certified SLPs, and (d) had computer access with internet capabilities. SLPs were currently practicing in California, Iowa, Indiana, Massachusetts, Ohio, Pennsylvania, and Wisconsin.

Voice Recordings

Voice samples were recorded using the Computerized Speech Lab by PENTAX at a sampling rate of 44.1 kHz (24 bits). A mouth-to-microphone distance of 2 in. was utilized for all recordings. Participants were asked to read or repeat the six sentences tested in the companion study. Stimuli are presented in Table 1. Additionally, a sustained /a/ vowel (5 s) and a conversational speech sample were recorded. Conversational speech samples were elicited using the prompts presented in Table 1.

Rating Procedures

SLPs completed two sessions of voice ratings approximately 2–3 months apart. SLPs were divided into two groups using a random number generator. Group 1 rated voices using VASs in Session 1 and Likert scales in Session 2. For Group 2, Session 1 ratings were made using Likert scales, and Session 2 utilized VASs. Voice rating sessions were identical, except for scale type utilized. In each session, SLPs were presented with 102 unique voices, with 23 (22.5%) voices presented twice in order to allow for the calculation of intrarater reliability. Voices were presented in randomized order (produced by random number generator) for Session 1. For Session 2, voices were divided into two groups: 50% of voices presented first, and another 50% presented second in Round 1. Both groups were rerandomized, and group order was reversed for Session 2, in order to counteract potential order effects while still maintaining randomized presentation order. Additionally, 10 voices (not the same repeated for interrater reliability) were presented a second time with limited stimuli (further details below), in

Table 1. Voice stimuli included in voice samples.

Standardized sentences	Sentence rationale	Sustained vowel	Conversational prompts (one of the following)
You and bob are eating.	Co-articulation of /a, i, u/	/a/ 5 s	Tell me how to make a peanut butter and jelly sandwich.
Harry has a hat.	Soft glottal attacks (voiced to voiceless transitions)		
We were away.	All voiced		
We eat eggs.	Hard glottal attacks		Tell me about your family.
Mama made muffins.	Nasal consonants		
Pet the puppy.	Oral pressure sounds		

order to assess the effect of stimuli type. A random number generator was used to select which voices would be repeated. Audio files were shared with raters in uncompressed file format to ensure recording fidelity. Raters listened to voice samples in quiet rooms using noise reduction headphones, which were shipped to SLP raters prior to study commencement (Pro-10, Universal Product Number 709466429447, OneOdio USA).

Across both sessions, raters were asked to maintain consistent volume when listening to voices. Prior to study commencement, raters were presented with two voices: one of a louder voice and one of a softer voice. Raters were asked to choose a listening volume that would work for both voices and maintained this volume for the remainder of the session. Raters were allowed to play audio clips more than once but were unable to change voice ratings once they were made. Sessions could be completed across multiple sittings if raters were experiencing fatigue or schedule constraints.

For each voice, six standardized sentences, a sustained /a/ vowel, and a conversational speech sample were presented. Conversational speech samples were limited to 25 s in order to mimic clinic conditions where voicing may be limited. As mentioned above, 10 voices were presented a second time with standardized sentences alone. This was done in order to conduct a preliminary investigation examining whether sentences alone might be sufficient for a reliable auditory-perceptual assessment of voice. As voice quality is known to change as a function of age and sex across childhood, raters were given each child's age and sex when ratings were performed. They were blinded to all other information, including voice status. Voice ratings were submitted online using the platform Qualtrics. VAS and Likert scales employed are described below.

After each voice was presented to raters, SLPs were asked if there was anything about the audio recording that prevented the auditory-perceptual assessment of voice. If five or more raters ($\geq 50\%$) indicated that recording quality was not sufficient for evaluation, a voice sample was eliminated from the analysis. Voices were not excluded for complaints regarding loudness or clip length, as these are ecologically valid variations in pediatric auditory-perceptual voice assessment.

VASs

The VAS was modeled after the scales currently employed in the CAPE-V (Kempster et al., 2009). VASs were anchored on the left with the term *normal* (0) and with the term *severely disordered* on the right (100). Raters were asked to place slider buttons along the VAS to the point they felt corresponded with the severity of overall voice quality, as well as individual parameters of voice quality, including breathiness, roughness, strain, pitch, and loudness. Possible values included 0 through 100, with all possible integers in between.

Likert Scale

The Likert scale was adapted from the GRBAS scale (Hirano, 1981). Options included *normal* (0), *mild* (1), *mild to moderate* (2), *moderate* (3), *moderate to severe* (4), and *severe* (5) dysphonia. *Mild to moderate* and *moderate to severe* options were included, as past study comparing the GRBAS and CAPE-V in children suggests that a rating of *moderate* on the GRBAS scale was used to encompass a wide variety of voice severity (Fujiki & Thibeault, 2021). In fact, a rating of *moderate* on the GRBAS has been documented to correspond with CAPE-V ratings ranging from 25 to over 70 (Fujiki & Thibeault, 2021). As such, this category was expanded in order to allow for more sensitive ratings of voices in the moderate range.

SLP Rater Feedback

Following both rating sessions, SLPs were asked to report on their experience with the VAS and Likert scales. This was done through an open-ended question asking SLPs about their experience with both scale types. SLPs were not required to answer this question; however, feedback was solicited twice. Responses were analyzed using qualitative content analysis. Responses were reviewed by two researchers and coded for themes line by line. Any disagreements were resolved in a consensus meeting.

Statistical Analysis

Descriptive statistics were calculated for all outcomes. The relationship between ratings made from the VAS and

Likert scales was assessed using Pearson's correlation coefficients (*R*). Additionally, mean VAS values corresponding with each Likert scale rating were calculated. Intrarater and interrater reliabilities were assessed using intraclass correlation coefficients (ICCs). This measure was chosen in order to facilitate comparisons with past investigations (Zraick et al., 2011). ICCs were calculated (two-way random effects) individually for both scale types. Specifically, we computed the ICC (2,*k*; Shrout & Fleiss, 1979), which informs how trustworthy an average rating is when raters are treated as random. ICCs were calculated separately for overall voice quality, breathiness, roughness, strain, pitch, and loudness. ICCs were also calculated across voice status, age group, and sex in order to illustrate the influence of these factors on rating reliability.

It was determined that a second analysis of interrater reliability was needed to further characterize agreement on the individual voice level and to clarify the contexts in which rating psychometrics varied significantly. This was deemed important because, although ICCs have been frequently used in past investigations, they do not allow for statistical comparisons of reliability across scale types or patient demographics. As such, Krippendorff's alphas were also calculated to assess interrater reliability and to investigate how individual parameters of voices impact rating psychometrics. Krippendorff's alpha assessed the level of consistency across multiple raters evaluating a common set of items. Possible values range from -1 to 1 , where 0 is indicative of no agreement and 1 represents complete agreement. Negative values indicate less agreement than would be expected by chance. Two Krippendorff's values were calculated for each voice: one for all ratings made using VASs and one for Likert scale ratings. In order to ensure that Krippendorff's alpha calculations were sufficiently powered, these values were calculated using ratings of overall voice quality as well as all individual parameters of voice. Multivariate regression was then used to identify factors that influence interrater reliability as measured by Krippendorff's alpha values. Factors in the model included voice status (voice disorder, vocally healthy), scale type (VAS, Likert), age (three groups; ≤ 6 , 7–11, and 12–17 years), and sex (male, female). Additionally, a Scale Type $\times$ Voice Status interaction was included. The α value was set at .01 for determining statistical significance in order to protect against Type 1 error. To illustrate the relationship between Likert and VAS ratings, a difference–mean plot adapted from Bland–Altman methodology was generated. Likert ratings were treated as numeric values consistent with common practice in auditory-perceptual voice research. Differences between paired VAS and Likert ratings were plotted against their mean to descriptively visualize systematic divergence and dispersion across severity levels. This analysis was intended to illustrate how discrete Likert categories map onto continuous VAS values.

Results

A total of 102 children were recruited for this study. Overall, 48 (47.1%) were female, and 54 (52.9%) were male. Sixty-one (59.9%) children were normophonic, and forty-one (40.1%) presented with diagnosed benign vocal fold lesions. The distribution of children across age, sex, and voice status is presented in Table 2. Mean VAS ratings are presented in Table 3.

ICCs

Interrater Reliability

Interrater reliability was high across all conditions. Values are presented in Table 4. Across both scales, ICCs were highest for overall voice quality (Likert ICC = .97, VAS ICC = .96) and lowest for pitch (Likert ICC = .82, VAS ICC = .78). ICCs were in excellent range across breathiness, roughness, and strain on both scales. Additionally, values remained high across all age groups, male and female children, and children with and without voice disorders.

Intrarater Reliability

Intrarater reliability was in moderate range for most raters (Koo & Li, 2016). When looking at individual raters on the Likert scale, ICC values ranged from .46 to .71. ICC values for the VAS ranged from .47 to .75. For three raters, ICC values were $> .10$ higher when using the Likert scale as opposed to the VAS. For two raters, the opposite was true, and ICCs for the VAS were $> .10$ higher. For the remaining five raters, ICC values were in roughly the same range across scales.

When comparing mean ICC values across parameters of voice quality, intrarater reliability was highest for overall voice quality (VAS = .56, Likert = .54) and breathiness (VAS = .57, Likert = .51). Values were lower for roughness, strain, and pitch (see Table 5). ICC values were similar across scale types for all the aforementioned parameters.

Table 2. Number of children across voice status, age groups, and sex.

Voice status	Age group (years)	Female	Male	Total
Vocally healthy	< 6	11	10	21
	7–10	12	8	20
	11–17	10	10	20
Voice disorders	< 6	7	8	15
	7–10	3	13	16
	11–17	5	5	10
All participants	< 6	18	18	36
	7–10	15	21	36
	11–17	15	15	30
	All ages	48	54	102

Table 3. Means and standard deviations for visual analog scale ratings.

Parameter	Overall	Voice status		Sex		Age (years)		
		Healthy voices	Voice disorder	Female	Male	< 6	7–10	11–17
Overall	18.98 (20.59)	11.86 (15.15)	28.66 (22.94)	16.53 (20.51)	21.25 (20.42)	21.63 (22.38)	22.08 (21.38)	11.28 (14.14)
Breathiness	13.18 (18.93)	7.02 (12.43)	21.55 (22.68)	12.08 (18.93)	14.19 (18.9)	15.48 (21.19)	15.12 (19.66)	7.52 (12.68)
Roughness	14.49 (17.96)	8.68 (12.72)	22.38 (20.81)	12.84 (17.98)	16.01 (17.82)	16.67 (19.44)	16.52 (19.13)	8.85 (12.19)
Strain	14.51 (19.28)	8.22 (13.39)	23.06 (22.51)	12.49 (18.46)	16.37 (19.85)	17.97 (20.75)	16.56 (20.62)	7.2 (12.22)
Pitch	3.37 (9.6)	2.53 (7.48)	4.51 (11.79)	2.62 (7.21)	4.07 (11.32)	2.18 (6.33)	4.86 (12.48)	2.84 (8.15)
Loudness	3.35 (10.69)	2.32 (8.21)	4.75 (13.21)	2.8 (9.21)	3.86 (11.87)	3.26 (10.23)	5.22 (13.79)	0.89 (3.34)

The largest difference between scale types was observed for loudness, where the ICC was .58 for the VAS and .44 for the Likert scale.

Krippendorff's Alphas

Mean Krippendorff's alpha values are presented in Table 6. Voice status was a significant predictor of interrater reliability ($\beta = .16$, $SE = 0.03$, $p < .001$). When controlling for scale type, sex, and age, interrater reliability was significantly greater for ratings of children presenting with voice disorders ($M = 0.33$, $SD = 0.29$) than for ratings of children with healthy voices ($M = 0.18$, $SD = 0.22$). Additionally, interrater reliability was also varied with age ($\beta = .01$, $SE = 0.03$, $p < .001$), as ratings of children in the < 6 group ($M = 0.33$, $SD = 0.31$) had significantly better interrater reliability than ratings of children in the 7 to 10 group ($M = 0.19$, $SD = 0.20$) or the 11 to 17 group ($M = 0.19$, $SD = 0.23$). No significant differences were observed across scale type ($\beta = .04$, $SE = 0.03$, $p = .113$) or sex ($\beta = .01$, $SE = 0.03$, $p = .702$).

Relationship Between Scales

Voice ratings made on the VAS and Likert scales were significantly correlated. The relationship between scales was strongest for overall voice quality ($r = .67$), followed by breathiness ($r = .64$), roughness ($r = .60$), and strain ($r = .61$). This relationship was weakest for pitch ($r = .39$) and loudness ($r = .33$). Table 7 contains the range of values from the VASs, which corresponds with values on the Likert scale. In summary, for overall voice quality, a rating of *normal* (0) on the Likert scale corresponded with an average rating of 5.2 ($SD = 9.1$) on the VAS. *Mild* (1) corresponded with a rating of 18 ($SD = 15.6$), *mild to moderate* (2) with 26.2 ($SD = 16.6$), *moderate* (3) with 36.2 ($SD = 20.5$), *moderate to severe* (4) with 38.0 ($SD = 26.1$), and *severe* (5) with 73.4 ($SD = 23.1$). The difference–mean plot revealed a strong, systematic relationship between magnitude of ratings and the difference between scale types, with differences increasing as overall severity increased (see Figure 1). This pattern reflects proportional scaling differences between the discrete Likert scale and the continuous

Table 4. Intraclass correlation coefficients for interrater reliability.

Parameter	Overall	Voice status		Sex		Age (years)		
		Healthy voices	Voice disorder	Female	Male	< 6	7–10	11–17
Visual analog scales								
Overall	.96 (.94, .97)	.91 (.86, .94)	.97 (.94, .98)	.96 (.94, .97)	.96 (.93, .97)	.96 (.93, .97)	.96 (.93, .97)	.93 (.89, .96)
Breathiness	.96 (.94, .97)	.90 (.86, .93)	.96 (.94, .97)	.97 (.95, .97)	.95 (.92, .96)	.96 (.93, .97)	.96 (.93, .97)	.95 (.92, .97)
Roughness	.95 (.92, .96)	.89 (.84, .92)	.95 (.92, .96)	.96 (.94, .97)	.93 (.89, .95)	.95 (.92, .97)	.93 (.89, .95)	.96 (.93, .97)
Strain	.94 (.91, .95)	.87 (.81, .91)	.95 (.91, .96)	.95 (.92, .96)	.93 (.89, .95)	.93 (.89, .95)	.94 (.89, .96)	.95 (.92, .97)
Pitch	.78 (.71, .83)	.51 (.32, .65)	.84 (.77, .89)	.45 (.22, .63)	.83 (.76, .88)	.48 (.23, .68)	.84 (.76, .90)	.51 (.23, .71)
Loudness	.92 (.89, .93)	.84 (.78, .89)	.94 (.91, .96)	.85 (.78, .89)	.94 (.91, .95)	.82 (.73, .89)	.94 (.91, .96)	.11 (–.39, .49)
Likert								
Overall	.97 (.95, .97)	.94 (.91, .96)	.96 (.94, .97)	.97 (.95, .98)	.96 (.94, .97)	.97 (.95, .98)	.97 (.94, .97)	.96 (.93, .97)
Breathiness	.96 (.94, .97)	.91 (.87, .94)	.96 (.93, .97)	.96 (.94, .97)	.95 (.94, .97)	.95 (.92, .97)	.96 (.94, .97)	.96 (.94, .97)
Roughness	.96 (.94, .96)	.92 (.89, .94)	.95 (.92, .96)	.97 (.95, .97)	.94 (.91, .96)	.96 (.94, .97)	.94 (.91, .96)	.95 (.92, .97)
Strain	.95 (.86, .96)	.90 (.86, .93)	.95 (.92, .96)	.96 (.92, .96)	.94 (.91, .96)	.96 (.92, .97)	.94 (.90, .96)	.93 (.88, .96)
Pitch	.82 (.76, .86)	.56 (.40, .96)	.87 (.91, .91)	.57 (.39, .71)	.86 (.80, .90)	.47 (.21, .66)	.89 (.82, .92)	.64 (.44, .79)
Loudness	.87 (.83, .90)	.80 (.72, .86)	.90 (.84, .93)	.77 (.68, .85)	.90 (.85, .93)	.79 (.68, .87)	.90 (.85, .93)	.27 (.11, .57)

Table 5. Intraclass correlation coefficients for intrarater reliability.

Scale type	Raters									
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10
Likert scale	.58 (.43, .65)	.46 (.29, .59)	.62 (.35, .74)	.61 (.42, .74)	.52 (.36, .65)	.56 (.42, .63)	.56 (.42, .62)	.71 (.53, .83)	.49 (.31, .60)	.61 (.51, .74)
VAS	.63 (.54, .71)	.47 (.36, .57)	.51 (.37, .62)	.48 (.31, .59)	.57 (.41, .73)	.66 (.45, .79)	.56 (.43, .68)	.50 (.38, .74)	.75 (.57, .89)	.68 (.53, .80)
Voice quality										
Scale type	Overall		Breathiness		Roughness		Strain	Pitch	Loudness	
Likert scale	.54 (.39, .67)		.51 (.36, .67)		.40 (.38, .51)		.42 (.32, .57)	.43 (.29, .54)	.44 (.30, .58)	
VAS	.56 (.40, .69)		.57 (.34, .68)		.47 (.32, .58)		.48 (.35, .60)	.43 (.29, .53)	.58 (.39, .65)	

Note. VAS = visual analog scale.

VAS rather than random disagreement. Visual inspection demonstrated the substantial dispersion of VAS values within each Likert category, particularly at moderate and higher severity levels, indicating that voices assigned the same Likert rating may vary on the VAS.

SLPs' Experience

Eight of 10 SLPs answered open-ended questions, and comments were similar across respondents. Four common themes were identified from SLPs' feedback. These included the following: (a) The Likert scale was faster than the VAS, (b) vocal performance sometimes varied across vocal tasks, (c) comorbid hypernasality made voice ratings more difficult, and (d) it can be difficult to determine what is "normal" in pediatric voices.

The Likert Scale Was Faster

SLPs indicated that the Likert scale was more efficient than the VAS. In the words of one SLP, "I preferred the Likert scale because it was so much faster to use."

Vocal Performance May Vary Across Voice Tasks

SLPs indicated that rating voices were more difficult when vocal performance varied across sustained vowels and connected speech. One SLP wrote, "I found myself second guessing myself a lot when there was variability between tasks—especially when the sustained vowel had roughness or pitch breaks not present in connected speech." Discrepancies in voice quality, pitch, and loudness were reported across sustained vowels and connected speech. Another SLP said, "It was hard to rate the sample and to

get a cumulative rating across all voice stimuli. For example, in some instances, the loudness was okay for sentences but soft for conversation."

Comorbid Hypernasality or Articulation Errors Made Voice Ratings More Difficult

SLPs reported that hypernasality made voice ratings more difficult. One SLP said, "Assessing patients/clients who had co-occurring articulation and/or resonance problems made the ratings a bit more challenging as I found myself trying to differentiate whether what I was hearing was truly voice or whether it was one of the other co-occurring abnormalities."

It Is Difficult to Determine What Is Typical for Pediatric Voices

Raters commented on the lack of data describing what constitutes typical voice quality in children. In the words of one SLP, "What I define as a 'normal' voice when rating pediatric voices is so much broader than with adult voices because children's voices are so much more variable." Another SLP commented that, "It can be difficult to determine what constitutes normal voicing. When I hear a little breathiness in a very young child's voice, it sounds 'normal,' but it is not absent breathiness."

Discussion

Pediatric auditory-perceptual assessments are an important aspect of voice evaluation in children (Fujiki & Thibeault, 2021). Unfortunately, the psychometrics of

Table 6. Means and standard deviations for Krippendorff's alphas measuring interrater reliability.

Scale type		Voice status		Sex		Age group (years)		
VAS	Likert	Healthy voices	Voice disorder	Female	Male	< 6	7–10	11–17
0.26 (0.27)	0.22 (0.25)	0.18 (0.22)	0.33 (0.29)	0.22 (0.28)	0.28 (0.24)	0.33 (0.31)	0.19 (0.20)	0.19 (0.23)

Note. VAS = visual analog scale.

Table 7. Visual analog scale means and standard deviations across Likert scale ratings.

Likert scale rating	Overall	Breathiness	Roughness	Strain	Pitch	Loudness
Normal	5.27 (9.1)	3.66 (9.2)	5.13 (10.1)	4.59 (9.3)	1.97 (6.9)	2.09 (8.2)
Mild	18.0 (15.6)	11.9 (14)	12.76 (12.5)	12.39 (15.2)	7.82 (11.9)	7.16 (13.2)
Mild to moderate	26.25 (16.6)	21.93 (17.0)	21.39 (15.9)	22.77 (16.7)	14.87 (14.5)	16.2 (19.3)
Moderate	36.26 (20.5)	28.87 (20.5)	28.81 (17.9)	33.42 (22.9)	12.75 (18.6)	14.18 (23.5)
Moderate to severe	38.03 (26.1)	38.58 (30.4)	38.56 (26.3)	39.51 (28.8)	65.8 (19.2)	31.71 (34.4)
Severe	73.44 (23.1)	56.97 (31.1)	74.76 (23.8)	55.78 (27.2)	51 (—)	—

Note. Em dashes indicate data not available.

auditory-perceptual voice ratings in children have eluded rigorous investigation. Additionally, no protocol for collecting pediatric auditory-perceptual voice ratings has been reported. As such, clinicians have been forced to adapt protocols designed for adults (Kelchner et al., 2010), and it is unclear how effective these modifications might be. The purpose of the current study was to develop the PAPE-V protocol, designed to promote reliable auditory-perceptual voice evaluations in children. As the initial stage of this endeavor, vocal stimuli designed specifically for children were developed, tested, and validated. The current study utilized these stimuli and examined rating procedures, scale type, and the manner in which demographic factors influenced rating reliability. Reliability was examined using complementary metrics that capture different aspects of agreement.

Interrater Reliability

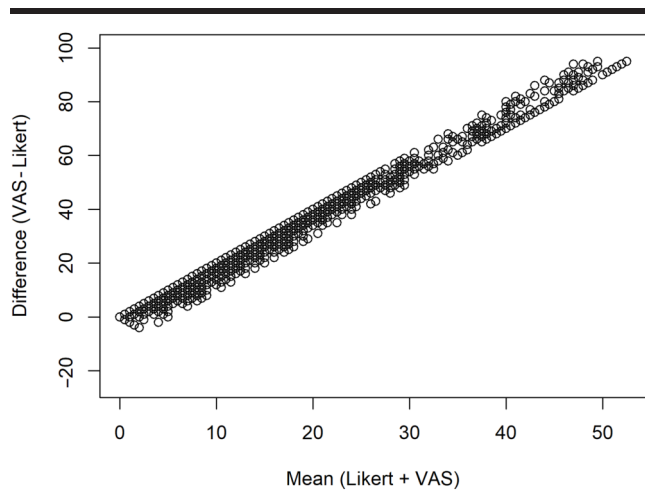
The VAS and Likert scales both demonstrated high interrater reliability as measured by ICCs. ICC values were in good to excellent range across all vocal parameters, scale types, and demographic subgroups. Values were highest for

overall voice quality and lowest for pitch, a pattern consistent with previous investigations of adult voices using the CAPE-V protocol (Zraick et al., 2011). Notably, ICC values observed in the current study were comparable to—or higher than—those previously reported in adults (Karnell et al., 2007; Nagle, 2025; Nemr et al., 2012; Solomon et al., 2011; Zraick et al., 2011). Although the relatively large number of voices included in the current investigation may have contributed to the stability of ICC estimates (Mehta et al., 2018), values remained high even when examining smaller subgroups defined by age, sex, and voice status. These findings indicate that raters were highly consistent in their relative judgments of pediatric voices; that is, they largely agreed on which voices were more or less dysphonic. This is particularly encouraging, given the absence of well-established norms defining typical voice quality in children (Fujiki, Venkatraman, & Heller Murray, 2025), and suggests that the PAPE-V protocol supports reliable auditory-perceptual voice assessment across a range of contexts.

Although ICC values were in good to excellent range, when interrater reliability was examined on the individual voice level, Krippendorff's alpha values were moderate, indicating that variability remained in how each voice was rated. This discrepancy was not unexpected, as ICCs and Krippendorff's alpha quantify different aspects of reliability. ICCs—particularly average-measure ICCs—reflect the consistency of raters' relative judgments and are tolerant of systematic differences in scale usage across raters. In contrast, Krippendorff's alpha is a more conservative estimate of interrater agreement that penalizes even small discrepancies in absolute ratings. Importantly, these metrics were also applied at different levels of analysis: ICCs summarized reliability across voices for each perceptual parameter, whereas Krippendorff's alpha captured agreement among raters for individual voices. The presence of high ICCs alongside lower alpha values suggests that raters were largely consistent in identifying which voices were more or less dysphonic but differed in the exact scale values assigned to specific children. This pattern likely reflects both the inherent variability of pediatric voice quality and the absence of well-established norms for defining typical voice in children.

Importantly, Krippendorff's alpha values suggested that reliability varied systematically as a function of voice

Figure 1. Difference–mean plot illustrating the relationship between Likert and VAS ratings of overall voice quality. Differences (VAS – Likert) are plotted against the mean of the two ratings. VAS = visual analog scale.



status and age. Voice ratings were significantly more reliable for children with voice disorders when compared to those who were vocally healthy. This may indicate that dysphonic voices present with more salient and consistently perceived deviations, whereas normophonic pediatric voices occupy a broader perceptual range. This finding may also reflect how little is known about what constitutes normophonic voice quality in children (Kreiman et al., 2021). Thus, it may be challenging—even for experienced SLPs—to determine what level of roughness, breathiness, or strain is typical for children. Preliminary evidence suggests that normophonic children may present with some level of breathiness or roughness (Aires et al., 2019), yet it is unclear at what threshold these qualities cause a voice to be perceived as dysphonic. It is likely the case that, in the absence of a voice complaint, a fair range of voice qualities may be considered typical for children.

Interrater reliability also varied with age. Surprisingly, interrater reliability was actually highest for the youngest age group (< 6 years). This was unexpected, given that the auditory-perceptual assessment of voice in young children is almost entirely unexplored. This finding may reflect the fact that voice quality is relatively stable between the ages of 4 and 6 years when compared to the ages of 7–10 or 11–17 years (Fujiki, Venkatraman, & Heller Murray, 2025; Heller Murray & Yucel, 2024). More variability may be expected in older children, and developmental changes may render voices more difficult to rate (Fujiki & Thibeault, 2025a; Lopes et al., 2013). For example, dramatic decreases in fundamental frequency occur in boys between the ages of 12 and 15 years (Infusino et al., 2015; Kent et al., 2021; Maturo et al., 2012). Additionally, changes in cepstral measures occur across childhood, with the most dramatic changes occurring after the age of 5 years (Heller Murray & Yucel, 2024). Many of these changes are related to the anatomical and physiological changes associated with puberty, which occur between the ages of 8 and 14 years in most children (Dorn & Biro, 2011; Kahane, 1982). Further studies are warranted to explore normative expectations of voice quality as children mature.

Intrater Reliability

Regarding intrater reliability, values were primarily in moderate range. Values fell within the range previously reported in adults using the CAPE-V, although maximum values skewed lower than those reported for ratings of adult voices for roughness and breathiness (Zraick et al., 2011). However, intrater reliability for strain exceeded that previously observed using the GRBAS scale in adults (De Bodt et al., 1997). These patterns support previous work suggesting that intrater reliability may be difficult to establish regardless of scale type (De Bodt et al., 1997;

Zraick et al., 2011). This finding may reflect the fact that raters had very limited information regarding the children whose voices they were rating. Given that pediatric voice quality may be quite variable, clinical voice ratings (particularly those of pitch and loudness) may be influenced by context. For example, raters may focus more heavily on breathiness, roughness, or strain in the presence of a voice complaint. Thus, rating voices without visual cues or medical history may have been particularly difficult; however, the audio-only design of the current study allowed for the standardization of stimuli and focused on voice quality alone. Clearly, more work characterizing normal voice quality in children is sorely needed.

Scale Type Considerations

Despite differences in scale format, no significant differences in interrater reliability were observed between VAS and Likert scale ratings using either ICCs or Krippendorff's alpha. This finding suggests that clinicians and researchers may have flexibility to choose which scale type they use with patients. This result differs from past work suggesting that VAS ratings may be more reliable than other common scale types (Baravieira et al., 2016; Kreiman et al., 2007; Martins et al., 2015; Yamasaki et al., 2017). In some cases, this may be partially attributable to the fact that VASs produce continuous data, whereas Likert or EAI scales produce ordinal data. Data types often examined using different tests of reliability (Gadermann et al., 2012). There is precedence for the current finding, however, as there are previous investigations reporting that Likert scales are similarly or more reliable than VASs (Karnell et al., 2007; Wuyts et al., 1999). Fortunately, the current data set suggests that either scale type may produce reliable voice ratings for children.

Considering that interrater reliability was in similar range across the VAS and Likert scales, it is important to consider SLPs' feedback regarding each scale type. SLPs overwhelmingly indicated that the Likert scale was faster to use. This finding reflects similar reports comparing clinician experience with the CAPE-V and GRBAS scales (Nemr et al., 2012). Given the demands of clinic schedules, SLPs may informally adapt protocols if there is not time to complete them as prescribed (Nagle et al., 2024). It may be that the ideal scale type is context dependent. For example, in busy clinic settings, the Likert scale may be more efficient and provide more salient categories. In research contexts, VAS ratings may be easier to analyze statistically and provide more fine-tuned measurements.

Regarding the relationship between scales, VAS and Likert ratings were moderately correlated for all parameters of voice quality. This finding supports past studies comparing the GRBAS and CAPE-V protocols in children and adults (Fujiki & Thibeault, 2021; Zraick et al., 2011),

which have reported that ratings of overall voice quality are most strongly correlated across scale types. If both VAS and Likert scales are utilized, the means from Table 7 could be used to determine what a rating on one scale might indicate on the other. Although this could not be done with perfect accuracy, it would allow for the general continuity of care between institutions. Similar data have been reported in adults to allow for conversion across scale types (Walden, 2025). It should be noted, however, that ratings of loudness and pitch were only weakly associated across scales, suggesting that these vocal parameters may require additional consideration.

Although either scale type may be utilized with similar reliability and findings will be correlated, it should be remembered that psychometric properties differ across scales. Visualization of the relationship between Likert and VAS ratings using a difference–mean plot highlighted important scale-dependent properties not captured by correlation alone. The strong linear trend observed reflects systematic differences in scale resolution, whereby VAS ratings expand perceptual severity at higher levels relative to the compressed categories of the Likert scale. This pattern does not indicate poor reliability but rather demonstrates that the two scales quantify perceptual severity differently and should not be considered interchangeable. Increasing dispersion of VAS values within Likert categories—particularly at moderate severity levels—suggests that, while Likert ratings may be efficient and clinically intuitive, they may mask variability captured by the VAS. This finding further supports the idea that scale selection should be context dependent. Likert scales may be advantageous in time-constrained clinical settings, while VASs may be preferable for research contexts requiring finer resolution and sensitivity to change.

Other Considerations

In children, it may be important to assess resonance (A. Lee et al., 2020). Evaluation of resonance is also heavily reliant on auditory-perceptual evaluation (Kuehn & Moller, 2000; Kummer, 2014). In developing an auditory-perceptual voice assessment protocol for children, it was tempting to add a category where raters might assess hypernasality or hyponasality. It was considered, however, that protocols for rating hypernasality in children already exist and have been the subject of considerable study (Chapman et al., 2016; John et al., 2006; Liu et al., 2022). In contrast to voice quality—which involves numerous components—hypernasality represents a single construct (nasal air leak) and can be rated based on the speech sounds on which it occurs (John et al., 2006). As such, it is recommended that if hypernasality needs to be assessed, a formally validated tool should be used, such as the Cleft Audit Protocol for Speech-Augmented-American Modification (Chapman et al.,

2016). Using a protocol specifically designed for the assessment of resonance will promote continuity between voice care and other settings in which SLPs assess resonance such as cleft palate and craniofacial anomaly clinics. The PAPE-V protocol can, however, include a space to note that further assessment is needed focusing on resonance.

SLPs noted that hypernasality and articulation errors made voice ratings more difficult. This is interesting considering that children in the current study presented with age-appropriate articulatory modifications. It is logical, however, that resonance and voice quality may not always be easy to separate perceptually. In fact, past study suggests that the presence of dysphonia or poor speech intelligibility weakens the relationship between auditory-perceptual and quantitative measurements of hypernasality (Fujiki et al., 2024). Thus, it is likely that the reverse could also be true and that hypernasality could influence ratings of dysphonia. Future work is needed to determine how voice quality can best be rated in children with resonance disorders.

It should also be noted that the current protocol was tested using SLP raters with expertise in the auditory-perceptual assessment of pediatric voices. As such, it is possible that rating reliability would be lower in raters with less familiarity either with this task or with the scale types (Franca et al., 2024). There is significant evidence to suggest that rater experience and even profession may influence voice rating psychometrics for adults (do Alves et al., 2024; Eadie & Kapsner-Smith, 2011; Helou et al., 2010; Kreiman et al., 1990). Although much of this work has examined rater populations who differed significantly in work experience, it is nonetheless possible that SLPs with less experience evaluating children might display lower rater reliability than more experienced clinicians. It is likely, however, that any protocol would require that SLPs familiarize themselves with pediatric voice. Future work should also determine how the PAPE-V functions across socioeconomic status and race, as these factors have been shown to influence how voice quality and other voice-related symptoms are perceived by clinicians and patients (Buckley et al., 2025; Fujiki & Thibeault, 2026; Sales et al., 2013).

Current findings suggest that the PAPE-V can be successfully used to conduct reliable pediatric auditory-perceptual evaluations. Task stimuli are appropriate for children, and rating reliability was high using both VASs and Likert-type scales. Thus, clinicians and researchers may choose whichever scale best suits their needs. VASs may be preferable for research contexts due to the ease with which these data can be analyzed statistically, as well as the broad range of rating options. The Likert scale may be more conducive for clinical contexts, as it is faster to

use, and the mild-to-severe values may be more meaningful clinically. Future work should clarify how voice rating reliability can be maintained in children with hypernasality or articulation errors.

Conclusions

Reliable auditory-perceptual evaluation of voice is feasible in children using the PAPE-V. This pediatric-focused auditory-perceptual voice protocol was developed in order to standardize and improve these evaluations in children. Findings indicate that the PAPE-V protocol constitutes an important step in promoting the reliability and efficiency in pediatric voice assessments. Future work should continue to refine this protocol to ensure that rating accuracy and reliability are maximized.

Data Availability Statement

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

Acknowledgments

This work was funded by an ASHFoundation 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), NIH Grant R01-DC004336 (awarded to Susan L. Thibeault), and National Center for Advancing Translational Sciences Grant K12TR004415 (awarded to Robert Brinton Fujiki). The authors thank Elizabeth Heller Murray, Maia Braden, Roseanne Clark, Loraine Syndey Kagan, Julia Leclerc, John Fredeking, Rita Patel, Christine Bueckers, and Brian Peterson for their expertise in performing voice ratings. They also thank Marie de Stadler, Abby Nimtz, John H. Munday, Kayla Beyer, Katlin Carpenter, Cora Ling, and Kate Gold Reilly for their roles in data collection.

References

- 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>
- Akin Şenkal, Ö., & Çiyiltepe, M. (2013). Effects of voice therapy in school-age children. *Journal of Voice*, 27(6), 787.e19–787.e25. <https://doi.org/10.1016/j.jvoice.2013.06.007>
- 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>
- Aronson, A. E., & Bless, D. (2009). *Clinical voice disorders* (4th ed.). Thieme.
- 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., Jetté, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clinical Linguistics & Phonetics*, 24(9), 742–758. <https://doi.org/10.3109/02699206.2010.492446>
- Baravieira, P. B., Brasolotto, A. G., Montagnoli, A. N., Silvério, K. C. A., Yamasaki, R., & Behlau, M. (2016). Auditory-perceptual evaluation of rough and breathy voices: Correspondence between analogical visual and numerical scale. *CoDAS*, 28(2), 163–167. <https://doi.org/10.1590/2317-1782/20162015098>
- Benoy, J. J., & Jayakumar, T. (2026). Effect of anchor voices and listener expertise on auditory-perceptual judgments of voice quality using the GRBAS scale. *Journal of Voice*, 40(3), 694–705. <https://doi.org/10.1016/j.jvoice.2023.12.011>
- Braden, M. N., & 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>
- Buckley, D. P., Dahl, K. L., Norotsky, R. L., Sauder, C., Tracy, L. F., & Stepp, C. E. (2025). Are auditory-perceptual evaluations of dysphonia by experienced voice clinicians affected by knowledge of speaker race? *American Journal of Speech-Language Pathology*, 34(3), 1341–1351. https://doi.org/10.1044/2025_AJSLP-24-00355
- Chapman, K. L., Baylis, A., Trost-Cardamone, J., Cordero, K. N., Dixon, A., Dobbeltsteyn, C., Thurmes, A., Wilson, K., Harding-Bell, A., Sweeney, T., Stoddard, G., & Sell, D. (2016). The Americleft Speech Project: A training and reliability study. *The Cleft Palate Craniofacial Journal*, 53(1), 93–108. <https://doi.org/10.1597/14-027>
- Choi, S. H., Yu, M., & Choi, C.-H. (2021). Comparisons of 4-point GRBAS, 7-point GRBAS, and CAPE-V for auditory perceptual evaluation of dysphonia. *Audiology and Speech Research*, 17(2), 206–219. <https://doi.org/10.21848/asr.200086>
- Cohen, W., & Wynne, D. M. (2015). Parent and child responses to the Pediatric Voice-Related Quality-of-Life Questionnaire. *Journal of Voice*, 29(3), 299–303. <https://doi.org/10.1016/j.jvoice.2014.08.004>
- 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
- De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Croux, C. (1997). Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*, 11(1), 74–80. [https://doi.org/10.1016/s0892-1997\(97\)80026-4](https://doi.org/10.1016/s0892-1997(97)80026-4)
- do Alves, J. 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(3), Article e20230175. <https://doi.org/10.1590/2317-1782/20232023175>
- Dorn, L. D., & Biro, F. M. (2011). Puberty and its measurement: A decade in review. *Journal of Research on Adolescence*, 21(1), 180–195. <https://doi.org/10.1111/j.1532-7795.2010.00722.x>
- Eadie, T. L., & Doyle, P. C. (2005). Classification of dysphonic voice: Acoustic and auditory-perceptual measures. *Journal of Voice*, 19(1), 1–14. <https://doi.org/10.1016/j.jvoice.2004.02.002>
- 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))
- 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>
- Franca, M. C., Boyer, V. E., & Tripathi, P. (2024). Effects of training in voice auditory-perceptual skills. *International Journal of Language & Communication Disorders*, 59(6), 2605–2616. <https://doi.org/10.1111/1460-6984.13103>
- 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., Johnson, R., Huang, S., & Thibeault, S. L. (2026). Pediatric Auditory-Perceptual Evaluation of Voice Part 1: Voice stimuli for children. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2026_JSLHR-25-00923
- Fujiki, R. B., Kostas, G., & Thibeault, S. L. (2024). Relationship between auditory-perceptual and objective measures of resonance in children with cleft palate: Effects of intelligibility and dysphonia. *The Cleft Palate Craniofacial Journal*, 61(8), 1245–1256. <https://doi.org/10.1177/10556656231162238>
- Fujiki, R. B., Sanders, P. W., Sivasankar, M. P., & Halum, S. (2022). Determining medical urgency of voice disorders using auditory-perceptual voice assessments performed by speech-language pathologists. *Annals of Otolaryngology, Rhinology & Laryngology*, 131(6), 579–586. <https://doi.org/10.1177/00034894211032779>
- 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., & Thibeault, S. L. (2026). Socioeconomic deprivation and vocal handicap in adults with voice disorders. *The Laryngoscope*. Advance online publication. <https://doi.org/10.1002/lary.70510>
- 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>
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, 17(1), Article 3. <https://doi.org/10.7275/n560-j767>
- 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>
- Hartnick, C. J., Rehbar, R., & Prasad, V. (2005). Development and maturation of the pediatric human vocal fold lamina propria. *The Laryngoscope*, 115(1), 4–15. <https://doi.org/10.1097/01.mlg.0000150685.54893.e9>
- 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))
- Hirano, M. (1981). Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice. In M. Hirano (Ed.), *Clinical examination of voice* (pp. 81–84). Springer.
- Hirano, M., Kurita, S., & Nakashima, T. (1983). Growth, development and aging of human vocal folds. In D. Bless & J. Abbs (Eds.), *Vocal fold physiology* (pp. 22–43). College-Hill Press.
- Hseu, A. F., Jo, S., Clark, R., Kagan, S., Kim, H.-Y., & Nuss, R. C. (2025). Dysphonia in preschool-aged children: Efficacy of voice therapy. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.02.010>
- Hseu, A. F., Spencer, G., Woodnorth, G., Kagan, S., Kawai, K., & Nuss, R. C. (2023). Barriers to voice therapy in dysphonic children. *Journal of Voice*, 37(3), 410–414. <https://doi.org/10.1016/j.jvoice.2021.01.008>
- 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>
- John, A., Sell, D., Sweeney, T., Harding-Bell, A., & Williams, A. (2006). The cleft audit protocol for speech—Augmented: A validated and reliable measure for auditing cleft speech. *The Cleft Palate Craniofacial Journal*, 43(3), 272–288. <https://doi.org/10.1597/04-141.1>
- Kahane, J. C. (1982). Growth of the human prepubertal and pubertal larynx. *Journal of Speech and Hearing Research*, 25(3), 446–455. <https://doi.org/10.1044/jshr.2503.446>
- Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice*, 21(5), 576–590. <https://doi.org/10.1016/j.jvoice.2006.05.001>
- Kelchner, L. N., Brehm, S., Weinrich, B., Middendorf, J., deAlarcon, 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. (1999). Improving the sensitivity and reliability of auditory-perceptual assessment. *Perspectives on Speech Science and Orolaryngology*, 9(1), 12–15. <https://doi.org/10.1044/ssod9.1.12>
- Kent, R. D., Eichhorn, J. T., & Vorperian, H. K. (2021). Acoustic parameters of voice in typically developing children ages 4–19 years. *International Journal of Pediatric Otorhinolaryngology*, 142, Article 110614. <https://doi.org/10.1016/j.ijporl.2021.110614>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- 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>
- Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research*, 33(1), 103–115. <https://doi.org/10.1044/jshr.3301.103>
- Kuehn, D. P., & Moller, K. T. (2000). Speech and language issues in the cleft palate population: The state of the art. *The Cleft Palate Craniofacial Journal*, 37(4), 1–35. https://doi.org/10.1597/1545-1569_2000_037_0348_saliit_2.3.co_2
- Kummer, A. W. (2014). Speech and resonance disorders related to cleft palate and velopharyngeal dysfunction: A guide to evaluation and treatment. *Perspectives on School-Based Issues*, 15(2), 57–74. <https://doi.org/10.1044/sbi15.2.57>
- Labaree, A., Bodt, M. D., & Nuffelen, G. V. (2024). Online versus classroom training of auditory-perceptual voice evaluation with the GRBAS-scale: A randomized controlled trial. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2024.09.046>
- 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.e13–6663.e21. <https://doi.org/10.1016/j.jvoice.2011.08.003>
- Lechien, 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>
- Lee, A., Potts, S., & Bressmann, T. (2020). Speech-language therapy students' auditory-perceptual judgements of simulated concurrent hypernasality and articulation disorders. *Clinical Linguistics & Phonetics*, 34(5), 479–492. <https://doi.org/10.1080/02699206.2019.1655666>
- Lee, Y., Lee, S., & Kim, J. (2025). Usefulness of direct magnitude estimation in auditory perceptual assessments for measuring dysphonia severity, roughness, and breathiness. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.03.031>
- Liu, Y., Lee, S. A. S., & Chen, W. (2022). The correlation between perceptual ratings and nasalance scores in resonance disorders: A systematic review. *Journal of Speech, Language, and Hearing Research*, 65(6), 2215–2234. https://doi.org/10.1044/2022_JSLHR-21-00588
- Lodhavia, A., & Kempster, G. B. (2026). Fidelity to the Consensus Auditory-Perceptual Analysis of Voice (CAPE-V): A pilot study. *Journal of Voice*, 40(3), 917.e1–917.e11. <https://doi.org/10.1016/j.jvoice.2023.12.009>
- Lopes, L. W., Lima, I. L. B., Azevedo, E. H. M., de Lima-Silva, M. F. B., Cavalcante, D. P., de Almeida, L. N. A., & de Almeida, A. A. F. (2013). Vocal characteristics during child development: Perceptual-auditory and acoustic data. *Folia Phoniatrica et Logopaedica*, 65(3), 143–147. <https://doi.org/10.1159/000355913>
- Martins, P. C., Couto, T. E., & Gama, A. C. C. (2015). Auditory-perceptual evaluation of the degree of vocal deviation: Correlation between the visual analogue scale and numerical scale. *CoDAS*, 27(3), 279–284. <https://doi.org/10.1590/2317-1782/20152014167>
- Maryn, Y., & Roy, N. (2012). Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *Jornal Da Sociedade Brasileira de Fonoaudiologia*, 24(2), 107–112. <https://doi.org/10.1590/S2179-64912012000200003>
- Maturo, S., Hill, C., Bunting, G., Ballif, C., Maurer, R., & Hartnick, C. (2012). Establishment of a normative pediatric acoustic database. *Archives of Otolaryngology—Head & Neck Surgery*, 138(10), 956–961. <https://doi.org/10.1001/2013.jamaoto.104>

- Mehta, S., Bastero-Caballero, R. F., Sun, Y., Zhu, R., Murphy, D. K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Statistics in Medicine*, 37(18), 2734–2752. <https://doi.org/10.1002/sim.7679>
- Nagle, K. F. (2025). Clinical use of the CAPE-V scales: Agreement, reliability and notes on voice quality. *Journal of Voice*, 39(3), 685–698. <https://doi.org/10.1016/j.jvoice.2022.11.014>
- Nagle, K. F., Helou, L. B., Solomon, N. P., & Eadie, T. L. (2014). Does the presence or location of graphic markers affect untrained listeners' ratings of severity of dysphonia? *Journal of Voice*, 28(4), 469–475. <https://doi.org/10.1016/j.jvoice.2013.12.011>
- Nagle, K. F., Kempster, G. B., & Solomon, N. P. (2024). Survey of voice-focused speech-language pathologists' usage of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2024.08.032>
- Nemr, K., Simoes-Zenari, M., Cordeiro, G., Tsuji, D., Ogawa, A., Ubrig, M., & Menezes, M. (2012). GRBAS and CAPE-V scales: High reliability and consensus when applied at different times. *Journal of Voice*, 26(6), 812.e.17–812.e.22. <https://doi.org/10.1016/j.jvoice.2012.03.005>
- Patel, R. R. (2023). Objective tools in pediatric voice therapy: A path to improved outcomes. *The Journal of the Acoustical Society of America*, 154(Suppl. 4), Article A336. <https://doi.org/10.1121/10.0023718>
- 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>
- Ricci-Maccarini, A., Schindler, A., Mozzanica, F., Fantini, M., Murry, T., & Dejonckere, P. (2022). Validity, reliability and reproducibility of the “extended GRBAS scale,” a comprehensive perceptual evaluation of dysphonia. *Journal of Voice*, 39(2), 393–402. <https://doi.org/10.1016/j.jvoice.2022.09.008>
- Rossi, N. A., Shen, B., Puckett, A. S., deSilva, B. W., & Wiet, G. J. (2026). Predictive efficacy of CAPE-V, pVHI, and CVA in detecting vocal fold pathology in children. *The Laryngoscope*, 136(3), 1500–150. <https://doi.org/10.1002/lary.70132>
- Salderay, Z. E., Yılmaz, M., Altınyay, Ş., Gölaç, H., & Gökdoğan, Ç. (2024). The effect of an indirect voice therapy approach on the voice of children with vocal fold nodules: A prospective cohort study. *Journal of Voice*, 38(4), 858–863. <https://doi.org/10.1016/j.jvoice.2022.01.001>
- Sales, N. J., Queiroz Gurgel, R., Rebelo Gonçalves, M. I., Côrtes Gama, A. C., Oliveira Cunha, E., Prado-Barreto, V. M., Todt-Neto, J. C., & Sampaio d'Ávila, J. (2013). Voice performance evaluation of street children from Aracaju, Brazil using perceptual–auditory and acoustic analyses. *Journal of Voice*, 27(5), 589–594. <https://doi.org/10.1016/j.jvoice.2013.02.009>
- Sauder, C. L., Kapsner-Smith, M., Simmons, E., Meyer, T., Doyle, P. C., & Eadie, T. L. (2024). The effect of rating method on reliability of judgments of strain across populations. *American Journal of Speech-Language Pathology*, 33(1), 393–405. https://doi.org/10.1044/2023_AJSLP-23-00174
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- 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>
- Tezcaner, C. Z., Ozgursoy, S. K., Sati, I., & Dursun, G. (2009). Changes after voice therapy in objective and subjective voice measurements of pediatric patients with vocal nodules. *European Archives of Oto-Rhino-Laryngology*, 266(12), 1923–1927. <https://doi.org/10.1007/s00405-009-1008-6>
- Tunç Songur, E. T., Hafizoğlu, M., Aydinli, F. E., İncebay, Ö., Parlak, M. M., & Balci, C. (2025). Analysis of the auditory-perceptual voice quality in older and younger adults without self-reported voice complaints. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2024.12.022>
- Walden, P. R. (2022). Perceptual Voice Qualities Database (PVQD): Database characteristics. *Journal of Voice*, 36(6), 875.e15–875.e23. <https://doi.org/10.1016/j.jvoice.2020.10.001>
- Walden, P. R. (2025). Severity cut-off ranges for auditory-perceptual evaluation using visual analog scales. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.01.018>
- Walden, P. R., & Rau, S. (2022). Individual voice dimensions' prediction of overall dysphonia severity on two auditory-perceptual scales. *Journal of Speech, Language, and Hearing Research*, 65(8), 2759–2777. https://doi.org/10.1044/2022_JSLHR-21-00689
- Wischhoff, O. P., Gouraram, V., Chumbley, T. J., Liu, B., & Jiang, J. J. (2025). Auditory-perceptual validation of acoustic chaos parameters in healthy and dysphonic voices. *Journal of Speech, Language, and Hearing Research*, 68(11), 5212–5225. https://doi.org/10.1044/2025_JSLHR-25-00155
- Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice*, 13(4), 508–517. [https://doi.org/10.1016/s0892-1997\(99\)80006-x](https://doi.org/10.1016/s0892-1997(99)80006-x)
- Yamasaki, R., Madazio, G., Leão, S. H. S., Padovani, M., Azevedo, R., & Behlau, M. (2017). Auditory-perceptual evaluation of normal and dysphonic voices using the Voice Deviation Scale. *Journal of Voice*, 31(1), 67–71. <https://doi.org/10.1016/j.jvoice.2016.01.004>
- 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))