Minimal Detectable Change of Experienced and Novice Listeners' Ratings of Overall Severity of Voice Quality

*Madeline Knutson, †Cara L. Sauder, ‡Jenny Vojtech, *,†Emily Wilson, *Lisa Zughni, and *,†Tanya L. Eadie,
*†Seattle, Washington, and ‡Boston, Massachusetts

SUMMARY: Objectives. The primary purpose of this study was to determine the minimal detectable change (MDC) scores—defined as the smallest amount of change greater than measurement error—for experienced and novice listeners' ratings of overall severity of voice quality for speakers with and without voice disorders. A secondary purpose was to determine the MDC scores as a function of speaker severity and listener experience.

Methods. Ten experienced listeners (ELs) and 10 novice listeners (NLs) rated overall severity of voice quality in 41 speakers with voice disorders and nine without voice disorders using a 100-mm visual analog scale across two rating sessions. MDCs at the 95% confidence level (MDC_{95}) were calculated. MDC_{95} scores also were calculated as a function of perceived speaker severity using averaged EL ratings from session one as the reference standard (categorized as typical, mild, moderate, and severe) for each listener group.

Results. The MDC_{95} for EL ratings of overall severity of voice quality for all speakers was 8.83 mm; for NLs, the MDC_{95} was 9.33 mm. As a function of speaker severity (n = 11 typical, n = 11 mild, n = 13 moderate, and n = 15 severe), MDC_{95} scores ranged from 3.70 to 15.18 mm for ELs and from 4.72 to 12.13 mm for NLs. For NLs, mildly dysphonic speakers required the largest MDC_{95} scores to be considered beyond measurement error. For ELs, MDC_{95} scores were largest for moderately dysphonic speakers.

Conclusions. Results have implications for measuring and interpreting auditory-perceptual outcomes for speakers with voice disorders when they are evaluated by experienced and novice listeners in both clinical and research contexts.

Key Words: Auditory-perceptual assessment—Listener experience—Dysphonia.

INTRODUCTION

Auditory-perceptual judgments are a fundamental component of voice evaluation.[1] Poor voice quality is one reason why many individuals with voice disorders seek voice treatment, and improved voice quality is often considered a primary outcome of treatment success. Yet, despite their importance, auditory-perceptual measures are subject to error, with much of the variability explained by task-related and listener-related factors.[1,2]

Several rating tools are used to perform auditory-perceptual assessment of voice in research and clinical settings. Two widely used tools include the Grade Roughness Breathiness Asthenia Strain (GRBAS) scale[3,4] and the *Consensus Auditory-Perceptual Evaluation of Voice* (CAPE-V).[5] The auditory-perceptual rating tool for evaluating voice disorders currently recommended by the *American Speech-Language-Hearing Association* is the CAPE-V.[5] The CAPE-V uses separate 100-mm visual analog scales (VASs) to capture overall severity and other parameters of voice quality: roughness, breathiness, and strain. Two separate scales are used to rate pitch and loudness. Clinicians use the VASs on the CAPE-V by marking the point on the 100-mm line that corresponds with the perceived severity of that parameter.[5] Scores are measured from the left endpoint (0 mm), representing what might be considered typical for that person's gender, age, and referent culture to the listener's mark (total indicated in mm); higher values represent worse voice quality.[5] For other voice parameters (breathiness, roughness, or strain), the left endpoint (0 mm) indicates none of that parameter, with higher values representing increased severity of that parameter.[5]

One of the difficulties in using rating scales like the CAPE-V to measure parameters of voice quality is related to a rater's reliability in using the scales for a given speaker or patient. When a clinician evaluates a speaker's voice, they compare that person's voice quality to the clinician's own internal standards for the voice quality dimension being rated. Internal standards are influenced by memory and experience, and may be unstable, even within a single rating session.[1] The comparison with internal standards may result in increased variability between listeners (ie, affecting *interrater reliability*),[1] or even variability for ratings of the same speaker by a single listener (ie, affecting *intrarater reliability*). Limitations with using these types of auditory-perceptual rating tools have been well documented in the literature.[1,2,6–9]

To address some of the difficulties with using rating scales for measuring voice quality, the original authors of the CAPE-V proposed several factors that might strengthen both the reliability and validity of the tool.[5] First, the authors proposed standard definitions for parameters such as overall severity, breathiness, roughness, and

Accepted for publication June 26, 2026.
From the *Department of Otolaryngology-Head and Neck Surgery, University of Washington, Seattle, Washington; †Department of Speech and Hearing Sciences, University of Washington, Seattle, Washington; and the ‡Department of Speech, Language, and Hearing Sciences, Boston University, Boston, Massachusetts.
Address correspondence and reprint requests to Madeline Knutson, Department of Otolaryngology-Head and Neck Surgery, University of Washington, 1959 NE Pacific Street, Seattle, Washington 98195. E-mail: mknuts@uw.edu
Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997
Published by Elsevier Inc. on behalf of The Voice Foundation.
https://doi.org/10.1016/j.jvoice.2026.06.049

strain.[5] This recommendation is based on research showing that even the simple provision of standardized definitions for voice parameters can improve rater reliability.[10] Second, the authors recommended use of continuous VASs for measuring vocal parameters; VASs may be more sensitive for detecting changes between and within speakers than ordinal scales, such as the 4-point ordinal (normal, mild, moderate, and severe) scale used in the GRBAS scale.[3,4] Finally, authors of the CAPE-V suggested that, in the future, exemplars for the voice parameters be considered for possible clinician training. The use of exemplars has previously been shown to improve rater reliability, ostensibly by circumventing the need to compare a given speaker's voice with an internal referent and instead comparing the speaker's voice with an external reference or anchor sample.[2,10,11]

Although a revised version of the CAPE-V[12] has recently been proposed to remove textual severity labels under each of the VASs and to modify several of the stimuli for evaluation, many sources of variability still exist. For example, the reliability of auditory-perceptual measures might vary with the experience of the clinician, as well as the severity of the speaker's dysphonia, and/or the parameter being rated, among other factors.[1,2,7,8,13,14] Given the strong reliance on auditory-perceptual judgments for managing the care of individuals with a variety of voice disorders, it is important that clinicians and researchers understand which factors have the potential to affect auditory-perceptual ratings so that they can validly use and interpret these measures across settings. These factors are further explored in the next section.

Factors affecting reliability of auditory-perceptual voice measures

A recent scoping review identified six broad factors that may influence rater reliability when clinicians or other listeners perform auditory-perceptual evaluation of voice disorders.[9] Specifically, attributes of the voice samples (eg, types of voice samples, severity of the voice samples), vocal parameters (eg, breathiness, roughness, etc), rater characteristics (eg, experience and professional background), training (eg, exposure to samples, training and feedback about ratings, etc), type of rating scale (eg, *n*-point or VAS, etc), and type of rating tasks were reported as the factors that most impacted listener reliability.[9] Although 101 studies were included in this scoping review, results and recommendations were based on 35 studies that conducted statistical tests to support their findings, suggesting many areas for future research.

The most studied factor (22%) among the 35 studies included in the scoping review was related to attributes of the voice samples.[9] Within that category, the authors noted that severity of the speech samples was an important factor that could affect rater reliability. According to Kreiman et al (1993) and others,[1,2,8] most listeners have considerable exposure to typical voices, so they have relatively stable internal standards by which to judge them. This is the reason why listeners usually demonstrate strong rater reliability when evaluating healthy or typical voices.[1] When evaluating speakers with dysphonia, listener responses are much more variable, in part because internal standards are affected by clinical experience, among other factors. However, unlike other speech disorders that may become more difficult to judge with increased severity, clinicians often are most reliable with severely dysphonic voices, while ratings of speakers with voice disorders in the midrange of the scale are often the most variable due to many factors (eg, perceptual ambiguity, measurement scale type, and differences in internal standards).[1,2,13–15] This result has important implications for clinical practice because most individuals with voice disorders fall into the midrange of voice quality severity at the time of their initial voice evaluations (ie, prior to voice therapy or other appropriate treatments).[16]

To better understand and interpret auditory-perceptual measures, clinicians must also consider how listener characteristics, such as the type or amount of experience, affect the consistency with which listeners make auditory-perceptual judgments, because listener reliability impacts the standard error of a measure.[8,17] With regard to intrarater reliability, most evidence has shown that experience is not strongly related to outcomes, particularly when listeners evaluate a parameter such as the overall severity of voice quality.[1,15,18] Of note, intrarater reliability of voice quality ratings is commonly assessed using a sampling of repeated stimuli (usually about 20–25%) within a given listening session in research studies.[1,8] Very few auditory-perceptual studies include repeated ratings of all samples[7,18,19]; this approach is more frequently observed in studies that assess the *test-retest reliability* of other types of measures, such as patient-reported outcomes.[20]

In their scoping review, Wang et al[9] suggested that increased rater experience was associated with stronger interrater reliability, similar to findings from two other studies.[9,15,21] However, there appear to be inconsistent results across studies.[1] For example, Kreiman et al (1993) reported interrater reliability values (intraclass correlation coefficients) ranging from 0.57 to 0.83, with no clear relationships between experience and reliability in their review of 10 studies that included listeners of different levels of experience. In fact, these authors posited that experienced listeners may *vary more* in their perceptual judgments because of their variable exposure to individuals with voice disorders as compared with novice listeners, who have primarily built their internal standards based on typical voices.[1] How experience relates to measurement error and our ability to interpret auditory-perceptual measures of voice quality clearly requires more study.

Finally, Wang et al[9] also reported that factors related to the rating task or type of rating scale could affect auditory-perceptual measures. Scale type may affect both the reliability and validity of such measures due to differences in the sensitivity of the scale as well as its ability to capture the construct.[1,9] For example, 100-mm VASs have increased

resolution when compared with traditional *n*-point rating scales; listeners are also at least as reliable when they make judgments on VASs when compared with ratings on *n*-point scales.[4,5,19,22] These are some of the reasons why VASs are part of the CAPE-V.[2,5,19]

One of the recommendations put forth by authors of the CAPE-V and its revised version, the CAPE-Vr, was that the speaker being evaluated should also be audio-recorded. This approach allows a clinician to more reliably compare a given patient's voice quality from pretreatment to posttreatment by comparing one voice to the other, thereby potentially bypassing unstable internal referents.[5,12] However, even using this approach for measuring changes over time, it is difficult to ascertain how much of a difference in rating scores might be needed on a perceptual scale to be confident that the change (or difference) goes beyond the measurement error of the tool. To determine whether differences in auditory-perceptual voice measures between or within speakers are beyond those simply due to measurement error, a minimally detectable change (MDC) score—the smallest amount of change greater than measurement error—must be determined.[23]

MDC and voice disorders

The MDC score provides confidence that either an observed change within a speaker or a difference between speakers is not simply due to random variability in the measure or variation that would be expected with multiple administrations of a tool.[23] Despite the importance of the MDC for determining whether a score on a particular tool is potentially clinically meaningful, this construct has only begun to be examined in the field of speech-language pathology, and in voice disorders specifically.[7,24–30]

Two recent studies have investigated MDCs for the auditory-perceptual evaluation of different types of voice disorders using 100-mm VASs.[7,27] Researchers in both studies calculated MDCs using 95% confidence intervals (MDC_{95}). First, Eadie et al[7] determined MDC_{95} scores for experienced and novice clinicians' ratings of overall severity of voice quality for 39 speakers with adductor laryngeal dystonia (ADLD) and 10 typical control speakers. Novice clinicians were those who had completed two graduate-level courses and a practical experience with voice disorders; experienced clinicians were those who reported 10 years clinical experience with voice disorders, on average. Clinicians in this study judged all speakers' voices for overall severity of voice quality using a 100-mm VAS (ie, similar to the CAPE-Vr; Kempster et al[12]); at least one week later, they rated all speakers again. Results showed that the MDC_{95} scores for all speakers were 11 mm for experienced clinicians and 9 mm for novices. In other words, when using a 100-mm VAS for overall severity, clinicians needed a rating that differed by at least 11 mm for that difference to be considered beyond measurement error. Interestingly, there did not appear to be a meaningful difference in MDC_{95} scores as a function of clinical experience.

Eadie et al[7] also categorized voice samples by perceived speaker severity (grouped as typical, mild, moderate, and severe). Results showed that MDC_{95} scores varied with the perceived severity of the speaker's voice quality, with higher scores for mild and moderate dysphonia and lower scores for speakers with typical voices and severe dysphonia. That is, MDC_{95} scores appeared to show similar patterns as those previously reported for rater reliability as a function of speaker severity, with MDC_{95} scores best (smallest) at the endpoints and most variable (biggest) in the midrange of the scale.[7,31]

In a second recent study, Smeltzer et al[27] investigated MDC_{95} scores for clinicians who rated pretreatment and posttreatment voice samples of 63 patients diagnosed with phonotraumatic vocal fold lesions who had either undergone voice therapy or laryngeal surgery and reported posttreatment voice improvements. In that study, experienced clinicians rated the pre- and posttreatment voice samples using all voice quality parameters of the CAPE-V,[5] with a subset of samples (19%) being rated a second time within a single session to calculate intrarater reliability. The results showed that the MDC_{95} scores were 15 mm for overall severity, 15 mm for roughness, 12 mm for breathiness, and 19 mm for strain. The authors suggested that the results were a good initial step for judging whether a "real" change beyond measurement error had occurred when using a scale such as the CAPE-V.[27]

One caution with extending findings from these two previous studies to a particular clinical or research setting is that MDCs are greatly affected by the context in which the values are obtained.[23] In other words, it is unknown whether results from one clinical population might extend to another or whether a clinician's experience might affect MDCs when using similar rating tools. For example, Eadie et al[7] assessed MDC_{95} scores for overall voice severity in symptomatic patients with ADLD that ranged from typical to severe. Authors of that study noted that the included ADLD speakers demonstrated predominantly strained voices. In contrast, Smeltzer et al[27] investigated MDC_{95} scores in patients diagnosed with phonotraumatic vocal fold lesions that included speakers with various amounts of breathiness, roughness, and strain. Other methodological differences, such as the type of reliability statistic used in the MDC calculation, can also affect results, in addition to characteristics of the sample under evaluation.[25] For example, Smeltzer et al[27] noted that because most speakers in their study fell into the mild-to-moderately severe range, the distribution of their sample could have affected their MDC_{95} values. They concluded that "future work should employ a cohort that involves speakers with voice qualities [that are] more evenly distributed across the severity continuum and investigate differences in MDCs… between severity groups" (p. 12).[27] However, no study has investigated MDCs for rating individuals with a variety of voice disorder etiologies across a wide range of voice severities for comparison with existing research.

Consequently, we sought to extend findings from these previous studies in two ways. First, we aimed to define

MDCs for two types of clinicians—experienced and novice—to further explore how experience might affect MDCs when using 100-mm VASs for rating speakers with a wide variety of voice disorders (ie, beyond ADLD or phonotraumatic lesions). Using a heterogeneous sample of individuals with varied voice disorder etiologies would allow for comparisons of MDCs to research studies using similarly heterogeneous samples. This type of sample is also representative of a typical voice practice, which often involves clinical evaluation of one patient after another with varied diagnoses. Second, given the known effects of speaker severity on rater reliability (and subsequently MDCs), we included speakers from a wide spectrum of dysphonia severities to determine how severity might also affect MDC scores. We selected overall severity of voice quality as the parameter for this study because we wanted to compare our results with these two previous studies, and because overall severity has strong face validity among speakers with a variety of voice disorder diagnoses, as well as both face and content validity among experts.[22] Thus, the primary purpose of this study was to determine MDC_{95} scores for both experienced listeners' (EL) and novice listeners' (NL) ratings of overall severity of voice quality in speakers with a wide variety of voice disorders and speakers without voice disorders using a 100-mm VAS. A secondary purpose was to determine MDC_{95} scores as a function of speaker severity and listener experience.

METHODS

This study was approved by the institutional review board (IRB approval No. STUDY00015518) at the University of Washington. All participants provided informed consent prior to completing study procedures.

Stimuli selection and preparation

This study included two groups of speakers. First, 41 adults (20 M, 21 F; mean age = 51.9 years; SD = 18.8; range = 25–85 years) with a variety of voice disorder diagnoses were selected from a database. The speakers were initially selected by three speech-language pathologist authors of this study (MK, CS, and TE) who were experienced in evaluating and treating individuals with voice disorders. The samples were selected to represent a range of voice severities, qualities, and diagnoses encountered in a clinical setting. Speakers who were perceived to have motor speech impairment (eg, imprecise articulation, velopharyngeal impairment) or who had confirmed neurological diagnoses (eg, Parkinson's disease, laryngeal dystonia) beyond vocal fold paralysis or paresis were excluded. The second group of speakers included 9 control speakers (4 F, 5 M; mean age = 45 years; SD = 19.9; range = 17–75 years) who reported no voice complaints. The control speakers were similarly selected from an available voice database and were screened by the same three authors to fall within typical limits for their age and sex; these speakers broadly matched the age range for the speakers with voice disorders Table 1 summarizes the voice diagnoses and other characteristics of the speakers included in this study.

In this experiment, stimuli were the second sentence of the Rainbow Passage.[32] These stimuli were taken from a database of voice recordings obtained in prior studies in a quiet room using a headset condenser microphone (AKG C420) routed to a digital audio recorder (Tascam DAP1) and digitized at a sampling rate of 44100 Hz and 16 bits, as well as recordings from a commercially available voice database (Disordered Voice Database Model 4337 [Ver. 1.03], 1994).[33] The second sentence, "The rainbow is a division of white light into many beautiful colors," was extracted for each speaker and peak-normalized for the listening procedure using sound-editing software. Judgments of the second sentence of the Rainbow Passage relate to listener judgments of the entire passage[32]; as a result, the second sentence is often used in both clinical and research studies to reduce listener burden.[2] Further, this sentence was recently added as a potential stimulus for standard

TABLE 1.
Speaker Demographics for Samples Included in Listening Procedure Including Average Age, Sex, and Number of Voice Disorder Diagnoses

Diagnosis	Avg Age (Years) (SD)	Sex (M, F)	Number in Sample
Benign vocal fold lesions	43.3 (16.7)	2 M, 11 F	11
Nonphonotraumatic vocal hyperfunction	57.7 (20.1)	2 M, 1 F	3
Unilateral vocal fold paresis/paralysis	52.5 (15.9)	7 M, 4 F	11
Bilateral vocal fold paresis	79.3 (5.1)	2 M, 1 F	3
Laryngeal cancer	67.7 (6.8)	2 M, 1 F	3
Papilloma	66.0 (18.4)	2 M, 0 F	2
Other (chronic laryngitis, laryngopharyngeal reflux, Reinke's edema, glottic stenosis, subglottic stenosis, unilateral vocal fold hemorrhage, and multiple diagnoses)	41.3 (18.0)	3 M, 5 F	8
Typical speaker	45.0 (19.9)	5 M, 4 F	9
Total	50.7 (19.0)	25 M, 25 F	50

voice evaluation in the CAPE-Vr.[12] All samples were entered into a custom software program (http://www.rubyonrails.org) that randomly generates speaker order, presents perceptual rating scales, and records listeners' responses.

Participants

Experienced listeners (ELs)

Ten speech-language pathologists (eight females, one male, and one nonbinary individual; mean age = 36.1 years; SD = 8.1 years) participated as ELs in this study. Nine (90%) self-identified as White, and one as Asian (10%), and none identified as Hispanic or Latino. ELs reported, on average, 10.7 years of experience (range = 3–24 years) in voice evaluation and treatment of individuals with voice disorders. Although there is no consensus in defining what constitutes an experienced listener, previous studies have included participants with at least two years of experience assessing and treating individuals with voice disorders.[1,7,18,19] All clinicians who participated in this study reported familiarity and practice using auditory-perceptual scales, including VASs, regularly for voice assessment, and none reported difficulties with hearing.

Novice listeners (NLs)

Ten (eight females, two nonbinary individuals; mean age = 25 years; SD = 2.4 years) individuals participated as NLs in this study. Individuals were recruited from the graduate student population in medical speech-language pathology at the University of Washington to serve as NLs. Six (60%) self-identified as White, two (20%) as Asian, and two (20%) as multiracial. One (10%) person identified as Hispanic or Latino, and nine (90%) did not identify as Hispanic or Latino.

All NLs had completed two graduate-level courses in voice disorders.[7] Coursework included practice with auditory-perceptual measures of voice and experience using VASs, knowledge of different types of voice disorder etiologies, and how these etiologies could affect vocal physiology and voice quality. Most participants had at least one practical experience with patients with voice disorders, although none had yet completed intensive voice-focused externships as part of their graduate program. All NLs passed bilateral hearing screening with thresholds at or below 20 dB HL for the octave frequencies between 250 and 4000 Hz.

Listening procedures

Both groups of listeners participated in two rating sessions using a test-retest paradigm. Sessions were held at minimum 7 days apart (EL mean = 15 days, range: 7–28 days; NL mean = 9.5 days, range: 7–14 days) to control for possible learning effects and to provide a robust sample for calculating test-retest agreement, which is required for calculating MDCs. Listeners were first oriented to the rating scale and then were provided a definition of overall severity of dysphonia as "a comprehensive measure of how 'good' or 'poor' the voice sample is judged to be."[34] Listeners were then presented four speech samples (two male and two female speakers) that represented a range of voice severities, from normal/typical to severe, to familiarize them with the range of severities they would hear in the rating procedure, and to remind them to use the entire range of the scale, if appropriate. None of the speakers used in this task were included in the experimental rating procedure. No additional training or anchor samples were provided to either rater group in the subsequent listening task.

Listeners rated all of the presented stimuli for overall severity of voice quality using a horizontally oriented 100-mm VAS, similar to the CAPE-Vr.[12] The endpoints of the scale were labeled as 0 = typical/healthy at the extreme left and 100 = severely dysphonic at the extreme right. Stimuli were presented in random order via a custom software program (http://www.rubyonrails.org) at a comfortable loudness over Sony MDR-7506 headphones (Sony Corporation) in a quiet environment. Similar to a clinical environment, listeners could play each sample as many times as they wished before making their ratings.[5] They controlled the pace of the stimulus presentation and could take breaks, as needed, to reduce fatigue. Each of the rating sessions took about 30 minutes.

Data analysis

All statistics were conducted using *SPSS* (Version 31).

Descriptive statistics

Descriptive statistics of overall severity ratings, including the mean, range, standard deviation, and standard error of the mean, were first calculated for each speaker and for each listener group within each rating session.[7] Each speaker's rating was based on averaged ratings of overall severity of voice quality for all speakers within each listener group, within each rating session (N = 50 samples x 10 EL ratings per sample = 500 ratings for ELs for each rating session; N = 50 samples x 10 NL ratings per sample = 500 ratings for the NLs for each rating session). Each speaker's score were based on averaged severity ratings across all raters in each listener group in an effort to control between-listener variability that could potentially contribute to between-speaker error in the ratings, and to ensure that results might generalize to a "typical" or representative listener in each listener group.[8,35]

Interrater reliability

To assess interrater reliability, we calculated intraclass correlation coefficients (ICCs). This calculation was performed to contextualize the amount of variance between listeners that could potentially affect the *between-speaker variance* (which is important for calculating MDCs). The ICC estimates were based on a mean rating (k = 10),

absolute agreement, and two-way random-effects model. Interrater reliability was excellent for NL ratings of overall severity of voice quality for all speakers in sessions 1 and 2: ICC(2,10) = 0.98 for both sessions. Interrater reliability also was excellent for EL ratings of all speakers: ICC(2,10) = 0.98 for both sessions 1 and 2.

MDC

The calculation of MDC requires consideration of the difference between scores on a measure collected at two rating sessions where a speaker's performance has not changed (ie, indicated in a metric of test-retest reliability) and the confidence level around which the MDC is determined.[23,28] Before calculating the MDC scores, the distribution of the differences between mean ratings for sessions 1 and 2 was examined for each group of listeners; both sets of data were visually inspected and Shapiro-Wilk tests were nonsignificant, meeting assumptions of normality.

MDC_{95} scores were first calculated from ratings of all 50 speech samples obtained from the two groups of listeners during rating sessions 1 and 2. MDC_{95} scores were calculated for averaged ratings of overall severity of voice quality for all speakers within each listener group. The MDC_{95} was calculated as: $MDC_{95} = 1.96 \times \sqrt{2} \times$ standard error of the measurement (SEM), where 1.96 is the z-score for the 95% confidence interval, and the square root of two accounts for the error across the two rating sessions.[23] The SEM, which is reported in the same units as the metric of interest (ie, mm), was calculated as: $SEM = SD_{pooled} \times \sqrt{(1-ICC)}$, where SD_{pooled} was the pooled standard deviation of all ratings from listening sessions 1 and 2 for each group of listeners.[36] Test-retest reliability was calculated using ICCs, based on a single measure ($k = 1$), absolute agreement, and two-way random-effects model at the 95% confidence interval for each group of listeners.[37] We selected the single-measure model because we were interested in how our results generalized to a clinician's rating from a single trial, as would be typical in clinical practice; agreement measures were selected to ensure we captured both systematic and random errors within the SEM; the random-effects model, which considers trial (or rating session) to be a random effect, was selected because we wanted to evaluate the test-rest reliability of our measure for use by other clinicians.[20]

To determine the effect of speaker severity on MDC_{95} and to permit comparisons between the listener groups, samples were divided into four severity groups based on the average EL severity ratings for each speaker in the first rating session using the following categories: typical = 0–9 mm, mild = 10–29 mm, moderate = 30–59 mm, and severe = 60–100 mm.[4] Before the samples were categorized, we first explored whether there were any systematic biases between the rater groups. An independent t test was used to compare mean ratings for all speakers obtained from each listener group at the 95% confidence level; no significant difference was found ($t(98) = -0.152$, $P > 0.05$), suggesting there was not a systematic bias in how each group evaluated the same set of speakers (ie, one group did not seem to perceive speakers as more severe than others, on average). As a result, we used averaged ratings across 10 ELs in the first rating session as the reference standard with which to categorize the severity of the voice samples. This approach has been used in other research areas where no uniformly agreed-upon objective measure of severity, including cutoffs between the categories of mild, moderate, and severe, exists.[35] MDC_{95} scores were then calculated for each speaker severity category for each listener group (ELs and NLs). Note that because only a single MDC_{95} is calculated for the group of speakers as a whole for each severity group and each listener group, only descriptive statistics are possible when examining differences between groups.

RESULTS

Test-retest agreement for ratings of overall severity of voice quality performed during listening sessions 1 and 2 was excellent for both ELs and NLs across all speakers (n = 50), with ICC(2,1) = 0.99 for NLs, and ICC(2,1) = 0.99 for ELs.

For both listener groups, test-retest agreement varied in strength as a function of the perceived severity of the speaker's voice quality. Perceived speaker severity was categorized using the average EL ratings from session 1 as the reference standard ($n = 11$ typical, $n = 11$ mild, $n = 13$ moderate, and $n = 15$ severe). Specifically, for NLs, ICCs were moderate in strength for mildly dysphonic speakers (ICC(2,1) = 0.55), good for typical speakers (ICC(2,1) = 0.73), and excellent for speakers with moderate (ICC(2,1) = 0.94) and severe dysphonia (ICC(2,1) = 0.95). For ELs, ICCs were good for typical (ICC(2,1) = 0.74) and moderately dysphonic speakers (ICC(2,1) = 0.71), and excellent for speakers with mild (ICC(2,1) = 0.86) and severe dysphonia (ICC(2,1) = 0.99).[38,39] The mean MDC_{95} for ratings of overall severity of voice quality for all speakers ($n = 50$) was 9.33 mm for NLs and 8.83 mm for ELs (Table 2).

When categorized by perceived speaker severity using the average EL ratings from session 1 as the reference standard, the MDC_{95} for ratings of overall severity of voice quality was lowest for speakers with perceived quality that was typical (MDC_{95} for NLs = 4.73 mm; MDC_{95} for ELs = 4.44 mm; Table 3). The highest MDC_{95} for ELs was 15.18 mm for moderately dysphonic speakers; for NLs, the highest MDC_{95} was 12.13 mm for mildly dysphonic speakers. Differences in MDC_{95} between ELs and NLs as a function of speaker severity ranged from less than 1 mm to 7 mm (Table 3 and Figure 1).

DISCUSSION

The purpose of this study was to determine MDC_{95} scores for ratings of overall severity of voice quality for speakers with a wide variety of voice disorders and those without

TABLE 2.
Test-Retest Agreement Values (ICCs) for Experienced and Novice Listeners' Ratings of Overall Severity of Voice Quality for All Speakers With and Without Voice Disorders, and for Speaker Severity Categories, Using Averaged Ratings From Experienced Listeners at Session 1 as the Reference Standard

Listener Group	All Speakers (N = 50)	Typical Voice Quality (n = 11)	Mildly Dysphonic (n = 11)	Moderately Dysphonic (n = 13)	Severely Dysphonic (n = 15)
Novice listeners	0.99	0.73	0.55	0.94	0.95
Experienced listeners	0.99	0.74	0.86	0.71	0.99

Note: Values are intraclass correlation coefficients (ICCs), based on a single rating (k =1), absolute agreement, and two-way random-effects model at the 95% confidence interval for each group of listeners.

voice disorders for experienced and novice clinicians. In addition to establishing MDC_{95} scores for each listener group, we sought to determine how these scores might differ as a function of a speaker's perceived overall severity of voice quality.

Overall, our findings indicated that regardless of a clinician's experience, speakers with and without voice disorders needed an overall severity of voice quality rating that differed by at least 9 mm on a 100-mm VAS for that difference to be considered real, and not due to random variability or error in using the tool. This is important information for individual clinicians to consider as they regularly use similar types of auditory-perceptual rating scales for judging voice quality in voice disorders.[2,5,12,22] Our results also complement the extant literature investigating MDCs for auditory-perceptual measures,[7,27] and self-reported measures in various voice populations.[29,30] For example, results from this study are consistent with a previous study that reported MDC_{95} scores of 11 mm and 9 mm for experienced and novice listeners who rated the overall severity of speakers with and without ADLD using a 100-mm VAS.[7]

In this study, we also sought to examine MDCs as a function of speaker severity because previous research has found that reliability for auditory-perceptual judgments of voice quality is best at the endpoints of a given scale, and most variable in the midrange; importantly, multiple

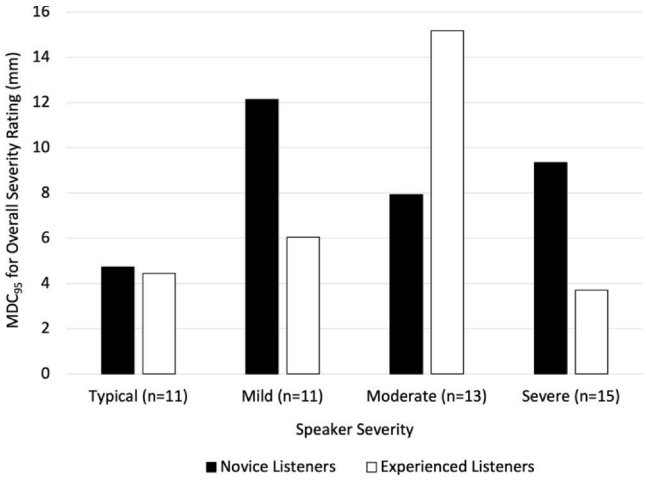

FIGURE 1. MDC_{95} scores for ratings of overall severity of voice quality on 100-mm VAS as a function of speaker severity and listener group (experienced and novice listeners).

sources of error associated with perceptions of speaker severity contribute to the standard error of a measure.[1,2,8,13–15] Consistent with one previous study (Eadie et al, 2025), we found that test-retest agreement and SEM estimates were more variable for mildly and moderately dysphonic speakers, and generally less variable for speakers at the endpoints, although test-retest agreement ranged

TABLE 3.
Experienced and Novice Listeners' Group Mean (*M*) Ratings of Overall Severity of Voice Quality From Sessions 1 and 2, Standard Deviations (*SD*), Standard Error of Measurement (*SEM*) Scores, and MDC_{95} Scores for Speaker Severity Categories, Using Averaged Ratings From Experienced Listeners at Session 1 as the Reference Standard

Listener Group	Samples Rated	*n*	*M* Session 1 (*SD*)	*M* Session 2 (*SD*)	*SEM*	MDC_{95}
Experienced Listeners	All speakers	50	39.39 (30.51)	39.86 (30.24)	3.19	8.83
Novice Listeners	All speakers	50	38.15 (29.44)	39.28 (29.62)	3.37	9.33
Experienced Listeners	Typical voice quality	11	4.20 (2.75)	4.77 (3.50)	1.60	4.22
Novice Listeners	Typical voice quality	11	5.67 (2.88)	5.69 (3.60)	1.71	4.73
Experienced Listeners	Mildly dysphonic	11	17.61 (5.88)	19.98 (5.89)	2.18	6.04
Novice Listeners	Mildly dysphonic	11	17.92 (5.42)	22.16 (7.51)	4.38	12.13
Experienced Listeners	Moderately dysphonic	13	42.90 (10.38)	41.23 (10.04)	5.48	15.18
Novice Listeners	Moderately dysphonic	13	39.31 (11.50)	38.09 (11.63)	2.86	7.92
Experienced Listeners	Severely dysphonic	15	78.11 (12.90)	78.97 (12.53)	1.33	3.70
Novice Listeners	Severely dysphonic	15	75.79 (14.98)	77.48 (14.96)	3.38	9.38

from moderate to excellent for both of our listener groups, regardless of the speakers' severity.[7] It should be noted that the ICC model used for test-retest reliability in this study was intentionally conservative and therefore represents lower-bound estimates. As hypothesized, across both listener groups, speakers with mild and moderate dysphonia needed larger MDC_{95} scores (EL = 6–15 mm; NL = 8–12 mm) than those with typical voices (EL = 4 mm; NL = 5 mm), for the difference to be considered beyond measurement error. Interestingly, novice listeners also needed differences of at least 9 mm when rating severely dysphonic voices for such ratings to be considered beyond measurement error. In contrast, experienced listeners needed differences of around 4 mm for ratings of severely dysphonic voices to be considered beyond measurement error. Thus, for experienced listeners, MDC_{95} scores were lowest at both endpoints of the rating scale, consistent with previous studies showing that ratings are most reliable, and contribute less to measurement error, for typical and severely dysphonic speakers.[7,31]

The MDC_{95} scores for ratings of overall severity for all speakers included in this study (about 9 mm for both ELs and NLs) were somewhat smaller than the MDC_{95} score of 15 mm recently reported by Smeltzer et al.[27] As previously discussed, MDCs are context-specific, and are affected by a number of factors, including the size of the sample and its severity range, among others. Differences in the characteristics of the included samples between these studies may explain some of the differences in results. For example, Smeltzer et al[27] investigated MDC values for expert listener ratings of all voice quality parameters, including overall severity, for 63 patients diagnosed with phonotraumatic vocal fold lesions who underwent voice therapy or laryngeal surgery and reported improvements. The average overall severity rating for their 63 rated speech samples (rated at both pretherapy and posttherapy) was about 19 mm (SD average = 17 mm; range: 0–99) on the CAPE-V.[27] In contrast, our study included speakers with a variety of vocal diagnoses, including those with benign laryngeal lesions, as well as other diagnoses. Importantly, our sample included a wide distribution of speaker severities, with an average overall severity rating of about 40 mm (SD average = 29 mm; range: 1–99 mm) across our sample, regardless of who performed the ratings. Given that we found larger MDCs (eg, up to 15 mm) for speakers who were moderately dysphonic in the current study, we would hypothesize that the MDC results from Smeltzer et al[27] were primarily driven by their lower average overall severity rating and limited standard deviation of their sample. These results are critical for clinicians to understand because most patients are mildly to moderately dysphonic when they receive their initial evaluation for voice therapy or are seen for other medical treatment.[16]

Interestingly, our study revealed no significant overall differences in the severity of ratings performed by novice and experienced clinician raters. However, MDC scores did appear to vary with speaker severity, with NLs demonstrating most difficulty in rating mildly dysphonic speakers and ELs demonstrating most difficulty in rating moderately dysphonic speakers (as demonstrated by lower ICCs and higher SEMs for both groups, respectively). Despite these differences between our listener groups as a function of speaker severity, none exceeded the overall MDC_{95} score of 9 mm, suggesting that differences were not meaningful and were within the measurement error of the scale. Previous research has shown inconsistent relationships between listener experience and auditory-perceptual judgments of voice, with some suggesting that experienced clinicians may agree more as a group,[9,21] and others finding no consistent differences.[2,18,40] This study adds to the literature suggesting that experience, at least as it was defined in this study, did not appear to meaningfully impact overall MDC scores for ratings of overall severity of voice quality in speakers with a variety of voice diagnoses.

Consistent with one previous study,[7] novice listeners included in this study were all from a single educational institution, where they had received similar training during their graduate education. The similarity in their training may have rendered these novice clinicians more similar than if they had been recruited from different institutions.[1,2] Further, although 7 of the 10 experienced listeners worked in the same healthcare setting, recent research suggests that even experienced clinicians may vary in how they use rating scales.[2] These factors warrant consideration when interpreting these results and how they might generalize to other clinicians who use these scales for evaluating voice disorders. Future studies might also investigate how additional clinical experiences such as voice-specialized clinical fellowships might also affect MDCs for clinicians over time.[15]

LIMITATIONS AND FUTURE DIRECTIONS

One factor that may affect a listener's reliability of auditory-perceptual judgments of voice relates to the dimension being evaluated. In this study, we sought to first establish MDCs for overall severity of voice quality because it has strong face and content validity across voice disorders.[22] Yet, most studies have demonstrated that overall severity of voice quality is the most reliable dimension assessed by listeners,[4,18,22,41] suggesting that the MDCs for overall severity of voice quality should not be extended to other voice parameters. This study utilized a sample of individuals with typical voices and voice disorders associated with a variety of etiologies that might have predominance in one or more voice parameters (eg, breathiness, roughness, and strain) that were not individually measured. Future studies should continue to establish MDCs for other auditory-perceptual parameters such as breathiness, roughness, and strain across a broad range of voice disorders and severities.[27] Beyond voice parameters, it is unknown whether MDCs should be established for particular voice diagnoses, as there is often considerable overlap between voice qualities that result from different voice diagnoses, with the exception of a few clinical populations (eg, ADLD). However, it is possible that results from such studies might help researchers interpret treatment outcome studies focused on patients with single clinical diagnoses (eg, treatments for vocal fold paralysis).

In this study, we also sought to minimize some sources of variability in our auditory-perceptual rating task by using a single sentence from the Rainbow Passage,[32] having clinicians perform evaluations based on research-quality recordings versus "live" speakers, and by using an undifferentiated visual analog scale without textual severity markers, similar to recent recommendations for a revised CAPE-Vr.[12] In addition, we used averaged ratings across ELs as our reference standard for categorizing our speaker samples into severities.[35] Future studies could also incorporate multiparametric objective measures that strongly relate to overall severity of voice, such as the Cepstral Spectral Index of Dysphonia or the Acoustic Voice Quality Index, to help corroborate these severities, although cutoffs for severity categories such as mild, moderate, and severe would need to be established.[42–44]

Although MDC scores provide information about the likelihood that differences on a particular tool reflect differences that are not simply due to measurement error, the size of the MDC score depends upon the test-retest metric that is selected and the confidence level around which the MDC is reported.[23] In this study, we chose to measure test-retest reliability using repeated ratings of all of the samples, rather than a sampling of repeated samples within a given session, strengthening the rigor and representativeness of our data. The difference in this approach may be one reason why some previous studies have sometimes shown poor-to-moderate intrarater reliability for speakers with a wide range of diagnoses,[22] even for overall voice severity. We also chose strict criteria for test-retest measures (ICCs, single measure, random effects, and agreement model) and a 95% confidence interval, similar to a previous study.[7] Yet, there are circumstances in which these strict criteria may not be necessary, for example, when determining the efficacy of a clinical intervention (eg, different voice therapy approaches) in which the consequences of a potentially incorrect decision are less of a risk than a scenario in which the consequences might be more dire (eg, likelihood of mortality from surgery), it might be sufficient to choose a 90% confidence interval, which decreases the size of the MDC.[45] Further, in treatment studies in which group-level data are investigated, it might be appropriate to use a test-retest measure that reflects average scores from a larger group of raters, which could further stabilize the ICC and SEMs, and decrease MDCs.[37] Whether the number of raters and/or speaker sample sizes might affect these measures needs further investigation, as these types of methodological decisions have the potential to affect outcomes. Similarly, the results of this study cannot be extended to other types of auditory-perceptual scale types (eg, GRBAS) that might be commonly used in clinical settings, including equal-appearing interval or ordinal scales. Future studies should investigate MDCs for these types of scales for direct comparison purposes.

Finally, MDCs in this study reflect the smallest actual change that exceeds typical measurement error based on listeners' ratings of the same voice samples twice to control for additional within-speaker variability that might be encountered in a clinical setting. Importantly, these scores do not reflect whether perceived differences might be clinically meaningful to individuals with voice disorders, their family members, or clinicians involved in their care: that is defined as the minimal clinically important difference (MCID).[46,47] A challenge in identifying the MCID of a metric is that it requires selection of an external anchor of meaningfulness, which is then compared against the tool under investigation. In their study, Smeltzer et al (2025) attempted to establish MCID thresholds for auditory-perceptual ratings of voice quality using clinician-rated anchors; they obtained an MCID threshold of 16.5 mm for overall severity. However, they also cautioned that these results might not extend to differences that might be perceived as meaningful to individuals with voice disorders or other communication partners.[27] Given previous research showing variable relationships between auditory-perceptual measures and other clinician- or patient-reported measures,[40,48] a multiparametric approach to establishing MCIDs may be warranted.[26]

CONCLUSIONS

Many factors affect the measurement error of auditory-perceptual rating tools, such as the CAPE-V[5] or CAPE-Vr,[12] which are heavily relied upon by clinicians who evaluate and treat individuals with voice disorders. We found that when clinicians, regardless of their experience level, rated the overall severity of voice quality of speakers with a variety of voice disorders and severities across the continuum, they needed a rating that differed by at least 9 mm on a 100-mm VAS for that difference to be considered beyond measurement error. Importantly, MDC scores also varied with the severity of the speaker's voice quality, with higher scores for those with mild and moderate dysphonia and lower scores for those with typical voices. Experienced listeners also tended to show lower MDC scores for ratings of those with severe dysphonia. Thus, regardless of experience, clinicians must remember that MDC_{95} scores should be interpreted with reference to the overall severity of voice quality of their patient, beyond the single MDC_{95} score for that parameter and type of rating scale often cited in research studies. Together with previous studies, these estimates are a step toward understanding how to interpret overall severity of voice quality outcomes in voice disorders, which are important for assessing and measuring treatment outcomes in a variety of clinical populations.

Data Availability

The auditory-perceptual data sets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Competing Interest

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

Acknowledgments

Thank you to members of the Vocal Function Lab for assistance with perceptual data analysis.

Note

Portions of this paper were presented at the annual Fall Voice Conference, October, 2024, Phoenix, AZ, and at the annual Convention of the American Speech-Language-Hearing Association, December 2024, Seattle, WA.

References

1. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality. *J Speech Lang Hear Res.* 1993;36:21–40. https://doi.org/10.1044/jshr.3601.21.
2. Nagle KF. Clinical Use of the CAPE-V scales: agreement, reliability and notes on voice quality. *J Voice.* 2022;39:685–698. https://doi.org/10.1016/j.jvoice.2022.11.014.
3. Hirano M. "GRBAS" scale for evaluating the hoarse voice & frequency range of phonation. *Clin Exam Voice.* 1981:83–84.
4. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice.* 2007;21:576–590. https://doi.org/10.1016/j.jvoice.2006.05.001.
5. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol.* 2009;18:124–132. https://doi.org/10.1044/1058-0360(2008/08-0017).
6. Barsties B, De Bodt M. *Assessment of Voice Quality: Current State-of-the-Art. Auris Nasus Larynx.* . Dublin, Ireland: Elsevier Ireland Ltd; 2015:183–188. https://doi.org/10.1016/j.anl.2014.11.001.
7. Eadie TL, Sauder CL, Marks KL, et al. Minimal detectable change of experienced and novice listeners' ratings of overall severity of voice quality in adductor laryngeal dystonia. *J Speech Lang Hear Res.* 2025;68:3119–3132. https://doi.org/10.1044/2025_JSLHR-24-00829.
8. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res.* 2005;48:323–335. https://doi.org/10.1044/1092-4388(2005/022).
9. Z. Wang Y. Wei W.S. Yu K.Y.S. Lee M.C.F. Tong T. Law Factors Influencing the Reliability of Auditory-Perceptual Evaluation of Voice: A Scoping Review Journal of Voice 2025 Elsevier Inc. doi: 10.1016/j.jvoice.2025.05.021.(Preprint posted online).
10. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice.* 2009;23:341–352. https://doi.org/10.1016/j.jvoice.2007.10.006.
11. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice.* 2006;20:527–544. https://doi.org/10.1016/j.jvoice.2005.08.007.
12. Kempster GB, Nagle KF, Solomon NP. Development and rationale for the consensus auditory-perceptual evaluation of voice—revised (CAPE-Vr). *J Voice.* 2025. https://doi.org/10.1016/j.jvoice.2025.01.022.(Published online).
13. Kapsner-Smith MR, Opuszynski A, Stepp CE, Eadie TL. The effect of visual sort and rate versus visual analog scales on the reliability of judgments of dysphonia. *J Speech Lang Hear Res.* 2021;64:1571–1580. https://doi.org/10.1044/2021_JSLHR-20-00623.
14. Sauder CL, Kapsner-Smith MR, Simmons E, Meyer T, Doyle PC, Eadie TL. The effect of rating method on reliability of judgments of strain across populations. *Am J Speech Lang Pathol.* 2024;33:393–405. https://doi.org/10.1044/2023_AJSLP-23-00174.
15. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. *J Speech Lang Hear Res.* 2011;54:430–447. https://doi.org/10.1044/1092-4388(2010/09-0205).
16. Fujiki RB, Thibeault SL. Examining relationships between GRBAS ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders. *J Voice.* 2023;37:390–397. https://doi.org/10.1016/j.jvoice.2021.02.007.
17. Stevens SS. Stevens G, ed. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects.* New York, NY: Wiley; 1975.
18. De Bodt MS, Wuyts FL, Van Dan Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice.* 1997;11:74–80.
19. Walden PR. Severity cut-off ranges for auditory-perceptual evaluation using visual analog scales. *J Voice.* 2025. https://doi.org/10.1016/j.jvoice.2025.01.018.(Published online).
20. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res.* 2005;19.
21. Helou LB, Solomon NP, Henry LR, Coppit GL, Howard RS, Stojadinovic A. The role of listener experience on consensus auditory-perceptual evaluation of voice (CAPE-V) ratings of post-thyroidectomy voice. *Am J Speech Lang Pathol.* 2010;19:248–258. https://doi.org/10.1044/1058-0360(2010/09-0012).
22. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V. *Am J Speech Lang Pathol.* 2011;20:14–22. https://doi.org/10.1044/1058-0360(2010/09-0105).
23. Riddle DL, Stratford PW. *Is This Change Real? Interpreting Patient Outcomes in Physical Therapy*. Philadelphia, Pennsylvania: F.A. Davis Compnay; 2013.
24. Barnett C, Green JR, Marzouqah R, et al. Reliability and validity of speech & pause measures during passage reading in ALS. *Amyotroph Lateral Scler Frontotemporal Degener.* 2020;21:42–50. https://doi.org/10.1080/21678421.2019.1697888.
25. Gates KE, Mefferd AS, Stipancic KL. Exploring methodological decisions for calculating the minimally detectable change in dysarthria: reliability, statistics, and standard error of measurement. *J Speech Lang Hear Res.* 2025;68:3771–3788. https://doi.org/10.1044/2025_JSLHR-24-00899.
26. Stipancic KL, Tjaden K. Minimally detectable change of speech intelligibility in speakers with multiple sclerosis and Parkinson's disease. *J Speech Lang Hear Res.* 2022;65:1858–1866. https://doi.org/10.1044/2022_JSLHR-21-00648.
27. Smeltzer JC, Stipancic KL, Toles LE. Minimal clinically important differences in CAPE-V auditory–perceptual ratings of voice quality. *J Speech Lang Hear Res.* 2025;68:2275–2290. https://doi.org/10.1044/2025_JSLHR-24-00503.
28. Stipancic KL, Yunusova Y, Berry JD, Green JR. Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *J Speech Lang Hear Res.* 2018;61:2757–2771. https://doi.org/10.1044/2018_JSLHR-S-17-0366.
29. Marks KL, Verdi A, Toles LE, et al. Psychometric analysis of an ecological vocal effort scale in individuals with and without vocal hyperfunction during activities of daily living. *Am J Speech Lang Pathol.* 2021;30:2589–2604. https://doi.org/10.1044/2021_AJSLP-21-00111.
30. Van Stan JH, Maffei M, Masson MLV, Mehta DD, Burns JA, Hillman RE. Self-ratings of vocal status in daily life: reliability and validity for patients with vocal hyperfunction and a normative group. *Am J Speech Lang Pathol.* 2017;26:1167–1177. https://doi.org/10.1044/2017_AJSLP-17-0031.
31. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am.* 2007;122:2354–2364. https://doi.org/10.1121/1.2770547.
32. Fairbanks G. *Voice and Articulation Drillbook*. 2nd ed., New York, NY: Harper & Row; 1960.

33. Disordered Voice Database Model 4337 Kay Elemetrics. Preprint posted online 1 1994 03.
34. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *J Acoust Soc Am.* 2002;112:3014–3021. https://doi.org/10.1121/1.1518983.
35. Mehta S, Bastero-Caballero RF, Sun Y, et al. Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Stat Med.* 2018;37:2734–2752. https://doi.org/10.1002/sim.7679.
36. De la Torre J, Marin J, Polo M, Marín JJ. Applying the minimal detectable change of a static and dynamic balance test using a portable stabilometric platform to individually assess patients with balance disorders. *Healthcare (Switzerland).* 2020;8:402. https://doi.org/10.3390/healthcare8040402.
37. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420–428. https://doi.org/10.1037/0033-2909.86.2.420.
38. Cicchetti DV, SSS. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic.* 1981;86:127–132.
39. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6:284–290.
40. Eadie TL, Kapsner M, Rosenzweig J, Waugh P, Hillel A, Merati A. The role of experience on judgments of dysphonia. *J Voice.* 2010;24:564–573. https://doi.org/10.1016/j.jvoice.2008.12.005.
41. Dahl KL, Weerathunge HR, Buckley DP, et al. Reliability and accuracy of expert auditory-perceptual evaluation of voice via telepractice platforms. *Am J Speech Lang Pathol.* 2021;30:2446–2455. https://doi.org/10.1044/2021_AJSLP-21-00091.
42. Lee JM, Roy N, Peterson E, Merrill RM. Comparison of two multiparameter acoustic indices of dysphonia severity: the acoustic voice quality index and cepstral spectral index of dysphonia. *J Voice.* 2018;32:515.e1–515.e13. https://doi.org/10.1016/j.jvoice.2017.06.012.
43. Maryn Y, Corthals P, Van Cauwenberge P, Roy N, De Bodt M. Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *J Voice.* 2010;24:540–555. https://doi.org/10.1016/j.jvoice.2008.12.014.
44. Peterson EA, Roy N, Awan SN, Merrill RM, Banks R, Tanner K. Toward validation of the cepstral spectral index of dysphonia (CSID) as an objective treatment outcomes measure. *J Voice.* 2013;27:401–410. https://doi.org/10.1016/j.jvoice.2013.04.002.
45. Donoghue D, Murphy A, Jennings A, et al. How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *J Rehabil Med.* 2009;41:343–346. https://doi.org/10.2340/16501977-0337.
46. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research how meaningful is it? *Pharmacoeconomics.* 2000;18:419–423. https://doi.org/10.2165/00019053-200018050-00001.
47. Jaeschke R, Singer J, Guyatt GH. Measurement of health status ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407–415. https://doi.org/10.1016/0197-2456(89)90005-6.
48. Behrman A, Sulica L, Tina. Factors predicting patient perception of dysphonia caused by benign vocal fold lesions. *Laryngoscope.* 2004;114:1693–1700.