This is the *Accepted Manuscript* of an article published by the American Speech-Language-Hearing Association (ASHA) in the *Journal of Speech, Language, and Hearing Research, 68*, 2275-2290 © 2025. The manuscript is reprinted here with permission from ASHA and is further available online https://doi.org/10.1044/2025_JSLHR-24-00503

Minimal Clinically Important Differences in CAPE-V Auditory-Perceptual Ratings of Voice Quality

Julianna C. Smeltzer[1,2], Kaila L. Stipancic[3], *Laura E. Toles[1]

[1]Department of Otolaryngology-Head and Neck, Voice Center, University of Texas Southwestern Medical Center, Dallas, Texas

[2]School of Behavioral and Brain Sciences, Department of Speech, Language, and Hearing, University of Texas at Dallas, Richardson, Texas

[3]Department of Communicative Disorders and Sciences, University at Buffalo, NY

*Corresponding Author:
Laura E. Toles
2001 Inwood Road
Dallas, TX 75390
214-645-2943
laura.toles@utsouthwestern.edu

Conflict of Interests: The authors have no conflicts of interests to disclose.

ABSTRACT

Purpose: This study aimed to determine the minimally detectable changes (MDC) and minimal clinically important differences (MCID) of auditory-perceptual ratings of voice quality using the CAPE-V scales (i.e., Overall Severity, Roughness, Breathiness, Strain).

Method: Participants (n= 63) included patients diagnosed with phonotraumatic vocal fold lesions who underwent either voice therapy or laryngeal surgery and reported post-treatment voice improvements. Nine expert voice-specialized speech-language pathologists rated the pre- and post-treatment voice samples using CAPE-V scales (i.e., via 100 mm visual analog scales with included textual labels for severity). Separately, raters judged the magnitude of perceived change between pre- and post-treatment samples using Jaeschke's Global Ratings of Change scale, which served as the anchor for MCID calculations. Intra-rater reliability and the standard error of measurement were used to calculate MDCs at the 95% confidence interval for each dimension. ROC curves were used to identify MCID thresholds, which were defined as values that optimized sensitivity and specificity while also exceeding the MDC.

Results: MDC values, representing thresholds for determining whether a true change has occurred, were 14.9 mm for Overall Severity, 14.6 mm for Roughness, 12.1 mm for Breathiness, and 18.7 mm for Strain. MCID thresholds, representing thresholds for determining clinically meaningful change, were 16.5 mm for Overall Severity, 16.5 mm for Roughness, and 15.5 mm for Breathiness. All potential MCID thresholds for Strain were smaller than the MDC value, thus a valid MCID threshold was not obtained.

Conclusion: This study represents the first known attempt to establish MDC and MCID thresholds for auditory-perceptual ratings of voice quality. The thresholds provide guidance for determining whether real and meaningful changes in voice quality have occurred in patients

undergoing treatment for phonotraumatic voice disorders. Future research should explore these values across various voice disorder populations and severity levels and incorporate patient-reported outcomes as anchors to enhance clinical decision-making and treatment outcomes in voice rehabilitation.

Key words: Voice Quality; Clinically Meaningful Change; CAPE-V

INTRODUCTION

Voice disorders occur in approximately 30% of individuals who live in the United States at some point in their lifetime (Roy et al., 2005) and are often accompanied by changes in voice quality (Carding et al., 2009). Classifying a patient's voice quality and severity through auditory-perceptual ratings is an important role of the speech-language pathologist (SLP) in the assessment and treatment of voice disorders (Roy et al., 2013). Voice quality ratings can serve as a marker for change to quantify the degree of improvement achieved throughout the course of voice therapy or following medical intervention for the voice problem.

Voice quality is multidimensional, resulting in complexity when attempting to perceptually discriminate dimensions of voice quality from one another. Though several auditory-perceptual scales existed prior to the Consensus Conference on Auditory-Perceptual Evaluation of Voice in June 2002, there was not a standardized protocol for rating voice quality. The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed by an expert consensus from the American Speech-Language Hearing Association Special Interest Group 3, Voice and Voice Disorders, to promote a best-practice approach to the evaluation and documentation of voice quality ratings (Kempster et al., 2009). The CAPE-V has become a widely-used, valuable clinical tool that includes a protocol for obtaining and rating voice samples using several distinct dimensions of voice quality.

Typical Use of the CAPE-V Scales

To evaluate the distinct auditory-perceptual dimensions of a voice using the CAPE-V, six perceptual features (i.e., overall severity, roughness, breathiness, strain, pitch, and loudness) are used (Kempster et al., 2009). Overall Severity is defined as an overall impression of the voice quality. Roughness is characterized by irregularity in the source of voicing, whereas Breathiness

is the audible air escape from voicing. Strain is defined as the perception of vocal effort or vocal hyperfunction when voicing. Pitch is conceptualized as the correlate to fundamental frequency, and Loudness represents the perceived sound intensity during voicing. Pitch and Loudness tend to be more straightforward to assess perceptually and have strong objective correlates (i.e., fundamental frequency and sound pressure level, respectively) compared to voice quality dimensions such as Roughness, Breathiness, and Strain which have more ambiguity in their relationship to objective metrics (Nagle, 2016). Thus, perceptual evaluation is currently the gold standard for measuring the severity of voice quality dimensions like Roughness, Breathiness, and Strain, in light of few direct objective correlates. Therefore, the current study focused primarily on the first four subscales of the CAPE-V to determine meaningful changes in voice quality dimensions.

Per CAPE-V protocol (Kempster et al., 2009), separate visual analog scales (VAS) consisting of 100 mm lines, are used to rate each voice dimension. The leftmost portion of the scale is intended to represent a voice that is unimpaired in that dimension, while the rightmost portion is intended to represent the most extreme impairment of each dimension. Three severity labels (i.e., Mild, Moderate, Severe) are provided on each scale to guide clinician ratings. In the original CAPE-V version developed during the consensus meeting in 2002, the scale used nonlinearly distributed severity labels. That is the version currently available for download on the ASHA website. Conversely, the version of the CAPE-V presented after peer review in the study by Kempster et al. (2009) uses linearly distributed textual markers. The clinician estimates the degree to which each voice quality dimension is present in the speech sample being rated by placing a tick mark on a scale position that corresponds to their percept of the feature, which is later translated into a number between 0 and 100 depending on the position marked. Numbers

closer to 0 represent a less deviant voice quality, whereas numbers closer to 100 represent a severely deviant voice quality. Clinicians are also given the opportunity to mark voice qualities as consistent (C) or intermittent (I) for each dimension. When using the CAPE-V to perform auditory perceptual ratings of a patient's voice, it is recommended to follow the CAPE-V procedures provided by the scale creators. The general procedure calls for the patient to sustain two vowels (i.e., /ɑ/ and /i/), read six sentences to elicit various laryngeal behaviors, and respond to a prompt to provide a sample of spontaneous speech, which are then rated using the CAPE-V scales. However, qualitative research has found variability with how the CAPE-V is implemented in clinical settings (Nagle, 2022). Clinicians do not always have patients progress through all recommended tasks due to time constraints during busy clinic sessions. In fact, a recent study by Lodhavia & Kempster (2024) found that few clinicians follow the standardized instructions for administration of the scale. For instance, clinicians may only rate voice quality on one of the recommended stimuli, on a different standardized prompt such as the Rainbow Passage, or on conversational speech. Additionally, many clinicians do not always or ever mark the C or I for the scales.

Reliability and Sources of Error in Auditory-Perceptual Ratings

Voice evaluation relies on psychophysical theory, which examines the relationship between stimulus characteristics and perceptual responses (Snook, 1999). In the context of auditory-perceptual analysis of voice, it is essential to understand how subjective judgments of voice quality align, or fail to align, with objective acoustic parameters of voice signals. Although the CAPE-V was designed to establish an accepted standard among clinician-rated voice severity and qualities, there is considerable variability associated with its use in everyday clinical practice which could introduce more room for error than objective correlates. Clinicians' auditory-

perceptual skills are influenced by a multitude of factors, including, but not limited to, the ways in which they perceive various speech features, how familiar they are with the speaker, and their clinical experience with rating voices (Kreiman & Gerratt, 2010; Kreiman et al., 1993; Kreiman et al., 1992; Nagle, 2022). Some studies have reported that there is adequate reliability and sensitivity of the CAPE-V (Eadie & Kapsner-Smith, 2011; Helou et al., 2010), whereas others show that some dimensions (e.g., Strain) have lower reliability than other features (Nemr et al., 2012; Zraick et al., 2011). Numerous strategies have been employed to address discrepancies in intra-rater reliability. These include using perceptual anchors (Eadie & Kapsner-Smith, 2011), single-variable comparison stimuli for matching strained voice quality (Park et al., 2023), matching tasks (Patel et al., 2012), and visual sort and rate methods (Granqvist, 2003; Sauder et al., 2024). Previous literature has also applied "signal detection theory" and the concept of "just noticeable difference" to understand perceptual thresholds in various sensory domains (Merfeld, 2011). These principles relate to a clinician's ability to detect subtle differences in voice quality, which are essential in distinguishing meaningful change from perceptual noise.

Both random and systematic errors can impact auditory-perceptual ratings of voice, which introduce variability in scores that are unrelated to actual changes in voice quality. Random errors include unpredictable influences such as fluctuations in a rater's attention, fatigue, and memory of previously presented stimuli, all of which can lead to inconsistent judgments (Poulton, 2023; Shrivastav et al., 2005). These random factors can add noise to perceptual judgments, making it challenging to detect true changes in a patient's voice. Systematic errors, often referred to as response bias or criterion errors, occur when raters apply the rating scale in a consistently biased way, such as using a narrow range of the scale or interpreting severity terms differently from others. For instance, a moderate rater might rate all

voices within a narrower portion of the scale, whereas a more liberal rater may use a broader range, leading to variability across raters (Shrivastav et al., 2005). Additionally, each rater's personal interpretation of the scale's severity range or the effect of previously rated samples can create systematic differences that increase error and make it more difficult to assess changes in voice quality. Non-stimulus factors, such as the use of perceptual anchors, training, and experience, can potentially control some of these typical sources of error. Eadie and Kapsner-Smith (2011) found that auditory anchors (i.e., consistent reference examples along the severity continuum) can reduce interrater variability by providing a standard against which all raters can calibrate their judgments. Acknowledging various sources of error in auditory-perceptual assessment is necessary as it impacts accuracy of ratings.

Identifying Change in Voice Quality

The primary objective of voice treatment is to facilitate meaningful change in a patient's voice. Due to the potential sources of error described above, variability of scores, and inconsistent reports of reliability, a major challenge that clinicians face when using the CAPE-V in a therapy setting is interpreting scores to accurately determine progress or changes in the severity of voice quality. There is currently insufficient literature describing what constitutes a true and meaningful change on the CAPE-V scales. This significantly limits clinicians' ability to use the CAPE-V as an outcome measure to identify changes occurring through treatment in a clinical setting or as a marker of change in research studies.

Interpreting change is not a problem that is limited to SLPs but is one that most healthcare professionals face. It is difficult to discern whether a change in scores on any scale/measure taken at different points in time is the result of a true change in patient status or simply due to measurement error. Given that there is an abundance of potential sources of error

192 in auditory-perceptual judgements, this is a significant challenge in the field of speech-language
193 pathology. For example, a clinician who rates their patient's voice quality at the initial visit
194 might find that their judgements of that patient's voice quality have improved by 10 mm on a
195 CAPE-V scale after a course of voice therapy, but it can be unclear whether this improvement
196 stems from effects of therapy or simply increased familiarity and adaptation to the patient. To
197 combat such challenges, there are two common concepts for assisting with interpretation of
198 change in clinical measures. The minimally detectable change (MDC) is defined as the smallest
199 change that indicates real change outside of measurement error (Beckerman et al., 2001; Haley &
200 Fragala-Pinkham, 2006). The MDC can provide a cutoff for whether real change has occurred,
201 but it does not necessarily reflect a significant *clinical* change (Beninato & Portney, 2011;
202 Stipancic et al., 2018). The minimal clinically important difference (MCID), a related but distinct
203 concept from the MDC, informs clinicians of the smallest degree of change that has clinical
204 significance (de Vet et al., 2006). Calculation of the MCID often requires an anchor or external
205 assessment to characterize the change as meaningful (Stipancic et al., 2018). To be useful and
206 meaningful, the MCID must be larger in magnitude than the MDC, because important change
207 cannot, theoretically, be within measurement error (Stratford & Riddle, 2012). MDC and MCID
208 have been established for a variety of measures in various fields within medicine and the
209 rehabilitation sciences (Amit et al., 2012; Horváth et al., 2017; Koopman et al., 2023; Salaffi et
210 al., 2004; Wright et al., 2017), but only recently have these concepts been applied in in the field
211 of speech-language pathology (Marks et al., 2021; Stipancic & Tjaden, 2022; Stipancic et al.,
212 2018; Young et al., 2018). Measures of clinically meaningful change are critically needed in the
213 voice disorders field, which is focused on rehabilitation and patient outcomes. Specifically,
214 investigating a clinically meaningful change in auditory-perceptual measures of voice quality

would greatly improve confidence in the clinical care of voice disorders, our ability to determine the magnitude of improvements associated with various treatment approaches, and the description of patient outcomes. Knowing what degree of change is true and clinically meaningful would help clinicians quantify the level of improvement in voice quality that patients have achieved as well as allow patients to observe objective changes through the course of their voice treatment.

Calculating clinically meaningful change values in voice quality requires voice samples in which voice changes have occurred (e.g., pre- and post-treatment measures, longitudinal data, etc.). The UT Southwestern Clinical Center for Voice Care holds a database that includes voice recordings of patients diagnosed with phonotrauma, both prior to and following treatment. Most individuals with phonotrauma will likely respond, to some degree, to voice therapy as it serves to reverse the maladaptive vocal hyperfunction associated with the disorder (Hillman et al., 2020; Holmberg et al., 2001). Other patients with phonotrauma will find voice improvements through surgical intervention (Childs, Rao, et al., 2022; Hillman et al., 2020). As a result, it would be expected that the CAPE-V scores for patients with this condition would improve following either behavioral voice treatment or surgical intervention. Using the pre- and post-treatment voice recordings from this database, the goals of this study were to determine (1) the magnitude of change in the voice quality dimensions of the CAPE-V scale that constitute real change (i.e., the MDC) and (2) the magnitude of change that represents a clinically important change in those ratings (i.e., the MCID).

METHOD

Participants

Speakers

This study was approved by the UT Southwestern Institutional Review Board (STU-2022-0655). Speech recordings of patients diagnosed with phonotrauma were gathered through a retrospective process using a database gathered for a separate project (Childs, D'Oto, et al., 2022; Childs, Rao, et al., 2022). Recordings from 63 patients were selected out of a total cohort of 124 patients that were screened. All patients had been diagnosed with a phonotraumatic voice disorder by one of three fellowship-trained laryngologists and had completed speech recordings prior to any intervention for their voice disorder and again after they noticed improvements in their voice status, as subjectively reported to their laryngologist during a post-treatment follow-up visit. Patients were chosen for inclusion if both speech recordings (i.e., pre- and post-treatment) were of good quality, at an adequate volume, without audio clipping, and without excessive background noise, as judged by two of the authors (J.C.S. and L.E.T.). Sixty-one patients were excluded based on these criteria. Of the 63 patients that were included in the ratings, 52 underwent a course of voice therapy only, and 11 underwent surgical removal of their phonotraumatic lesion(s). Patients were randomized into three equal groups (n = 21) for ratings based on the vocal severity in the pre-treatment recordings to ensure a relatively balanced severity level in each rating group and to control the length of the listening task. A voice-specialized speech-language pathologist with 13 years of experience categorically rated the overall severity of dysphonia of baseline recordings as either mild, moderate, or severe, solely for the purpose of creating balanced (i.e., in terms of voice impairment severity) groups of speakers. Each of the three groups contained ten patients who were rated as mild, eight who were

rated as moderate, and three who were rated as severe. Demographic information of patients who were included in each group can be seen in **Table 1**.

Table 1. Demographic information of all patients and within assigned groups for the listening task.

Parameter	All Patients	Group 1	Group 2	Group 3
Diagnosis *n*				
Nodules	25	9	7	9
Polyp	13	3	6	4
Cyst	3	2	0	1
Pseudocyst	22	7	8	7
Age *M(SD)*	35.7 (14.5)	35.1 (14.6)	35.8 (16.3)	36.3 (13.2)
Sex *n F/M*	54 F/9 M	19 F/2 M	19 F/2 M	16 F/5 M
Surgery *n*	11	3	4	4
Therapy *n*	52	18	17	17
Voice impairment severity *n*				
Mild	30	10	10	10
Moderate	24	8	8	8
Severe	9	3	3	3

Expert Raters

Nine voice-specialized SLPs were recruited to participate as expert raters. Per IRB protocol, verbal informed consent was obtained from raters. Raters had a mean of 13.7 years (range = 7 to 30 years, standard deviation = 7.8 years) of clinical experience, with voice

disorders comprising at least half of their current caseload. All raters were familiar with and had experience using the CAPE-V in their clinical practice. To minimize burden on individual raters, and to reduce the time required to rate samples, raters were randomized into three groups and each group rated one of three groups of recordings described above.

Procedures

Speaking Voice Samples

Patients were recorded while reading The Rainbow Passage in the clinic exam room during visits with their treating laryngologist at the UT Southwestern Clinical Center for Voice Care. Audio recordings were collected immediately prior to laryngoscopy exams. In the laryngology clinic, an omnidirectional condenser lapel microphone [Olympus ME-15] was placed on the patient's collar at a mouth-to-microphone distance of approximately 15 cm and 45° from the left oral angle. The microphone was connected to a VaultStream nCare [Olympus] medical recorder which contains a built-in pre-amplifier. All acoustic recordings were sampled at 44.1 kHz. Patients were instructed to read the Rainbow Passage in their typical, comfortable speaking voice. The Rainbow Passage was chosen as the rating stimuli because it was the only standardized speech task collected at laryngologist appointments consistently. Each patient was recorded at their initial evaluation with the laryngologist prior to initiating any treatment. All patients included in this study also recorded the Rainbow Passage as a post-treatment speaking voice sample at a follow-up visit, at which time they reported to the laryngologist that their voice status had improved. We did not include any patients who reported no change or worsening of their voice in order to reduce the amount of variability in this initial study of clinically meaningful change in voice features.

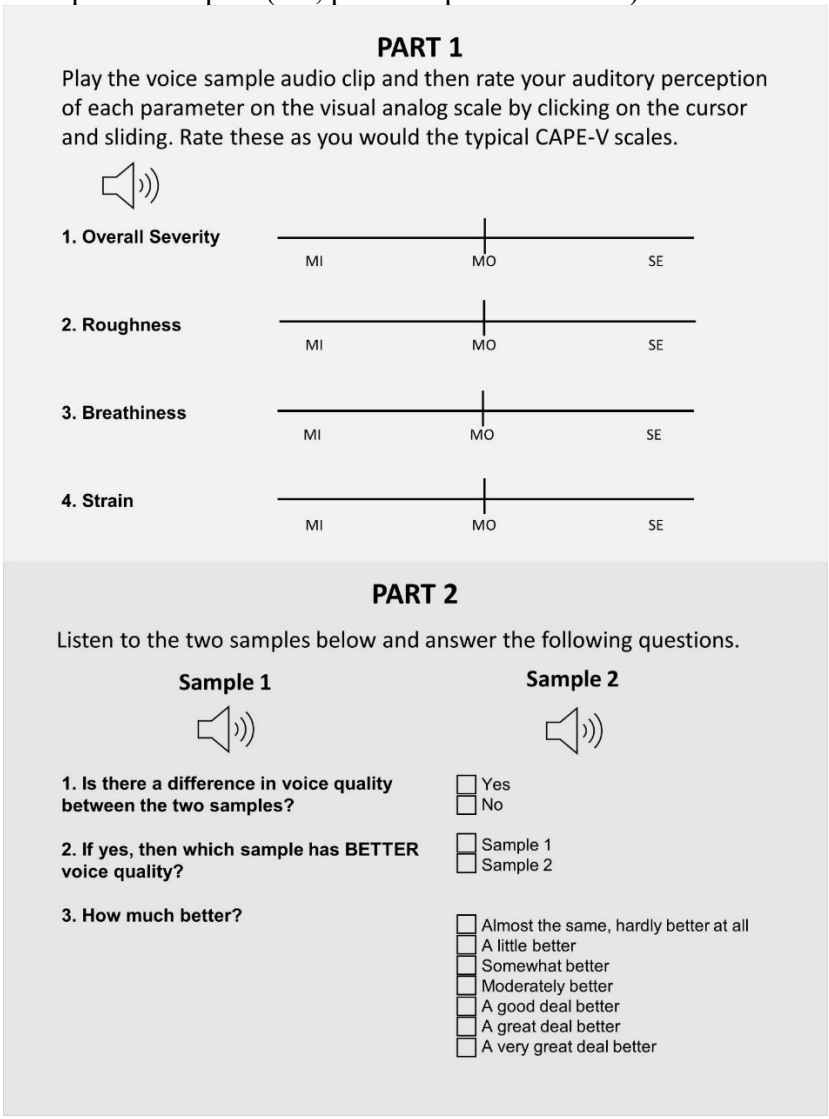
Because samples were collected in a clinical setting and were not originally intended for research, not all collected samples were appropriate for rating due to the quality of the recording. It is difficult to control gain during patient visits with the nCare equipment that was used to collect audio samples in the laryngology visits because the audio waveform is not visualized while the recording is being obtained. Therefore, samples with excessive background noise (e.g., from laryngoscopy equipment, talking in the background, etc.) or audio signals that were too soft or loud to be rated confidently were not included, per procedure described in the Participants section above.

Listening Tasks

Expert raters were sent a link through REDCap electronic data capture tools hosted at UT Southwestern (Harris et al., 2019; Harris et al., 2009) for completing their ratings. Raters were instructed to use a personal computer or laptop, to use high-fidelity over-ear headphones, to sit in a quiet room, and to ensure a comfortable listening volume. The use of noise canceling headphones was not specified. The REDCap survey presented tasks in two sections, which are described in detail below: one section for ratings of CAPE-V auditory-perceptual scales and a second section with the MCID task. For both rating tasks, all speech samples were randomized using Excel. Raters were presented audio recordings of sentences 2 and 3 from The Rainbow Passage (i.e., "The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon") from each speaker's pre- and post-treatment recordings. Therefore, each listener heard 42 unique speech samples. Instructions provided to the raters can be visualized in **Figure 1**.

Figure 1. Rendering of the tasks presented to each rater. Part 1 shows the auditory-perceptual rating tasks that raters completed for each speech sample (n = 42 unique speech samples and n = 8 repeated speech samples). Part 2 shows the Global Ratings of Change (GROC) questions that

were used to calculate the minimal clinically important differences. Raters were presented with two speech samples (i.e., pre- and post-treatment) from the same participant



Auditory-Perceptual Ratings. The first section of the listening task involved rating voice samples based on the four voice quality-related scales in the CAPE-V protocol: 1) Overall Severity, 2) Roughness, 3) Breathiness, and 4) Strain. Raters were presented with a recording for a single speaker at one time-point at the top of the page and were asked to rate the four scales below that. Each scale was presented as a 100 mm VAS in the form of a slider bar. The slider cursor was set at 50 mm. The rater had to click and hold the slider bar and move to the left or

right to rate the sample per their internal expert-based auditory-perceptual standards. Per the CAPE-V, word-based anchors were provided on each scale, with "MI" (i.e., mild) being near the left part of the scale, "MO" (i.e., moderate) being in the middle, and "SE" (i.e., severe) being toward the right part of the scale (Kempster et al., 2009). Raters were able to replay samples as many times as they needed before advancing to the subsequent sample, which is a common procedure for rating the CAPE-V (Nagle, 2022). Each group rated the randomized pre- and post-treatment speaking voice samples from the 21 patients in their respective group. Eight (19%) randomly selected speech samples in each group were rated twice to permit calculation of intra-rater reliability.

MCID Task. Methods for calculating the MCID were based on those completed by Stipancic et al. (2024). The MCID task was implemented to provide an external anchor to judge clinical meaningfulness between samples within individual patients (Jaeschke et al., 1989). We used Jaeschke's global ratings of change (GROC) to establish this anchor. Raters listened to both speaking voice samples from the same patient, one immediately following the other. The order of speaker and pre-/post-treatment sample combinations were randomized. As presented in **Figure 1**, they were then asked the following question: "Is there a difference in voice quality between the two samples?" and were given response options "yes" and "no." If they responded "no", they moved onto the next set of samples. If they responded "yes", they were then asked to select which of the two samples had the better voice quality using buttons to choose "Sample 1" or "Sample 2." Finally, they were asked, "how much better?" and were provided the following seven-point response options:

1. Almost the same, hardly any better at all
2. A little better

3. Somewhat better
4. Moderately better
5. A good deal better
6. A great deal better
7. A very great deal better

Data and Statistical Analyses

All data and statistical analyses were completed using R (R Core Team, 2013).

Inter-rater Reliability

Inter-rater reliability was calculated separately for each feature (i.e., Overall Severity, Breathiness, Roughness, and Strain) and separately for each of the three groups of listeners who heard the same list of stimuli, using intraclass correlation coefficients (ICCs; two-way models for agreement based on the average ratings of multiple raters). ICC values of less than 0.5 represent poor reliability, values between 0.5 and 0.75 represent moderate reliability, values between 0.75 and 0.9 represent good reliability, and values greater than 0.90 represent excellent reliability (Koo & Li, 2016).

Minimally Detectable Change

MDCs were calculated using the samples which were repeated for each rater (eight repeated samples per rater = 72 samples) to calculate intra-rater reliability. Intra-rater reliability was calculated separately for each feature using the same model of ICCs as above between the first CAPE-V rating completed by a rater and the second rating of the same sample completed by the same rater. ICCs were calculated using the 'irr' package in R (Gamer et al., 2022). Then, the standard error of measurement (SEM) for each feature was calculated with the following formula: SEM = SD x Ö(1-r), where the SD was the standard deviation of the ratings completed

by listeners (excluding the reliability trials). Then the MDCs at the 95% confidence interval were calculated with the following formula, used in prior literature (Fulk & Echternach, 2008; Haley & Fragala-Pinkham, 2006; Stratford & Riddle, 2012): $MDC_{95} = 1.96 \times Ö2 \times SEM$ separately for each feature.

Minimal Clinically Important Difference

MCIDs were calculated following methods in Stipancic et al. (2024); Stipancic et al. (2018). Receiver operating characteristic (ROC) curves were used to define thresholds that maximize sensitivity and specificity of a specific cut-off point on the CAPE-V scales for distinguishing between participants whom raters said experienced a change on the GROC scale. For example, a ROC curve was used to determine the diagnostic accuracy, sensitivity, and specificity of the Overall Severity scale on the CAPE-V for distinguishing between participants for whom raters rated the difference between their samples a 1 on the GROC scale (i.e., "Almost the same, hardly better at all") vs. participants for whom raters said there was no difference between their samples. Typically, a cut-point for identifying "clinically important" change would be selected on the anchor scale, which in this case, is the GROC scale. However, because this was the first investigation to use the GROC scale for voice features, we used a data-driven approach to identify the appropriate cut-point. Therefore, ROC curves were created for every point on the GROC scale for each of the four voice features. The point on each of the feature scales that maximized sensitivity and specificity, that was also larger than the MDC would ultimately be identified as the MCID.

RESULTS

Descriptive statistics, including mean, standard deviation, minimum, and maximum values, for pre-treatment and post-treatment ratings of each voice quality dimension for each

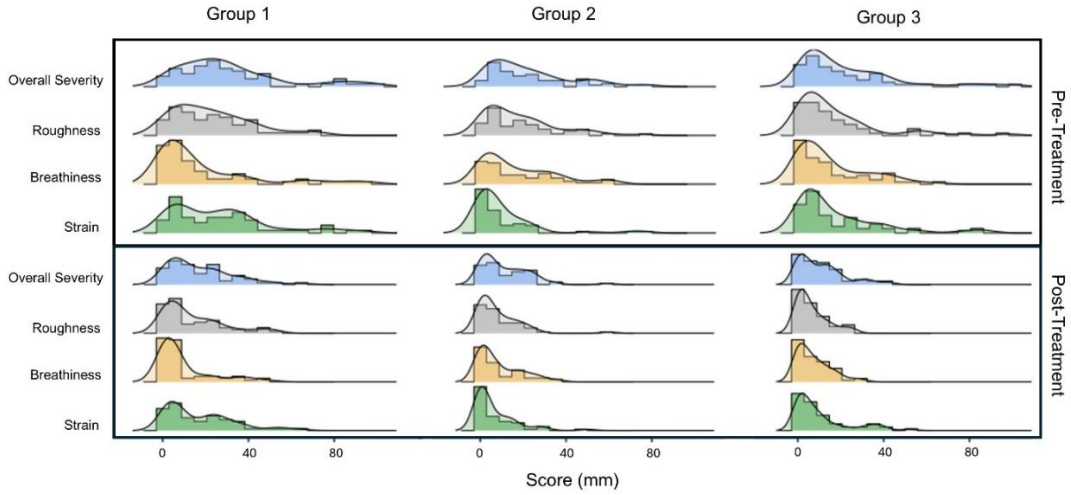
group of samples are provided in **Table 2** and can be visualized in **Figure 2**. Each voice quality dimension was present in the samples across each group, and average scores decreased post-treatment. Distributions were positively skewed, with more density of scores in the mild to moderate dysphonia ranges.

Table 2. Summary statistics of each CAPE-V voice quality dimension at each time point by group

			CAPE-V Scale			
Group	**Timepoint**	**Parameter**	**Overall Severity**	**Roughness**	**Breathiness**	**Strain**
1	Pre-Treatment	Mean (SD)	30.98 (24.72)	22.56 (18.08)	18.11 (24.18)	25.90 (22.61)
		Minimum	0	0	0	0
		Maximum	99	72	96	94
	Post-Treatment	Mean (SD)	17.16 (14.72)	13.86 (14.03)	9.16 (12.50)	16.78 (15.56)
		Minimum	0	0	0	0
		Maximum	65	52	50	64
2	Pre-Treatment	Mean (SD)	21.54 (18.94)	17.41 (16.68)	16.65 (17.11)	9.02 (14.82)
		Minimum	0	0	0	0
		Maximum	78	75	61	75
	Post-Treatment	Mean (SD)	11.43 (11.28)	7.97 (10.04)	8.44 (9.88)	6.37 (9.74)
		Minimum	0	0	0	0

		Maximum	59	60	36	50
3	Pre-Treatment	Mean (SD)	21.24 (22.65)	19.57 (6.35)	14.03 (15.79)	16.65 (20.12)
		Minimum	0	0	0	0
		Maximum	100	97	63	85
	Post-Treatment	Mean (SD)	10.79 (10.98)	6.35 (7.18)	7.13 (7.63)	10.27 (12.82)
		Minimum	0	0	0	0
		Maximum	43	25	30	52

Figure 2. Histograms with density curves for each voice quality dimension of the CAPE-V at each time point by group.



Inter-rater Reliability

Inter-rater reliability for each group of listeners for each of the four voice features are displayed in **Table 3**. Across groups, ICCs for Overall Severity ranged from 0.89 to 0.94, for Breathiness ranged from 0.78 to 0.92, and for Roughness from 0.82 to 0.92, indicating good to

410 excellent reliability for these three features. Across groups, ICCs for Strain ranged from 0.41 to
411 0.89, indicating that one of the listening groups had poor reliability for this feature whereas the
412 other groups had good reliability.

413

414 **Table 3.** Inter-rater reliability estimates using intraclass correlation coefficients with confidence
415 intervals in brackets for each voice feature across listener groups.

Listener Group	Overall Severity	Roughness	Breathiness	Strain
Group 1	0.906 [0.843-0.947] $p < .001$	0.817 [0.693-0.897] $p < .001$	0.917 [0.86-0.953] $p < .001$	0.785 [0.639-0.878] $p < .001$
Group 2	0.887 [0.813-0.935] $p < .001$	0.849 [0.749-0.913] $p < .001$	0.776 [0.629-0.871] $p < .001$	0.408 [0.018-0.659] $p = 0.021$
Group 3	0.942 [0.901-0.967] $p < .001$	0.917 [0.861-0.953] $p < .001$	0.876 [0.792-0.93] $p < .001$	0.892 [0.819-0.939] $p < .001$

416

417 **Minimally Detectable Change**

418 Intra-rater reliability, SEMs, and MDCs are presented in **Table 4**. MDCs across all
419 speaker participants were as follows: 14.91 mm for Overall Severity, 14.62 mm for Roughness,
420 12.06 mm for Breathiness, and 18.74 mm for Strain.

421

422

Table 4. Intra-rater reliability, standard error of measurements (SEM), and minimally detectable change (MDC) estimates (in mm) for each voice quality parameter.

Parameter	**Intra-rater reliability (Intraclass correlation coefficient)**	**SEM (SD x √(1-r))**	**MDC_{95} (1.96 x √2 x SEM)**
Overall Severity	0.90	5.38	14.91
Roughness	0.88	5.27	14.62
Breathiness	0.90	4.35	12.06
Strain	0.83	6.76	18.74

Minimal Clinically Important Difference

ROC curves for each of the four voice quality parameters are displayed in **Figure 3**. The left-uppermost corner of the ROC plots is the point that maximizes sensitivity and specificity and was defined as the MCID. The black diagonal line in each plot represents a diagnostic tool that works at chance (i.e., has no greater than a 50/50 chance of distinguishing between changed and unchanged participants). **Table 5** displays the results of the ROC analyses including AUCs, specificity, sensitivity, accuracy, and the threshold on the CAPE-V scales which maximizes sensitivity and specificity. Because, theoretically, the MCID must be larger than the MDC (i.e., clinically significant change cannot be within measurement error (Stratford & Riddle, 2012), thresholds identified by the ROC analyses that were larger than the calculated MDCs could be considered the MCID. As can be seen in **Table 5**, a valid MCID was not obtained for the Strain subscale because all thresholds were smaller in magnitude than the MDC.

Figure 3. Receiver operating characteristic curves for (A) Overall Severity, (B) Roughness, (C) Breathiness, and (D) Strain, across all levels of the Global Ratings of Change Scale.

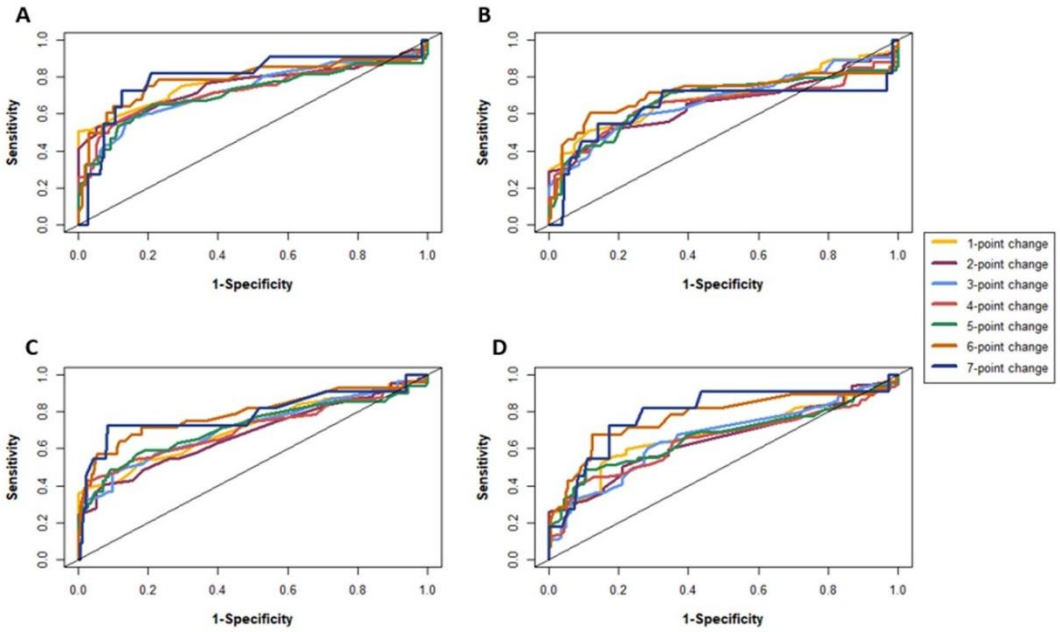


Table 5. Receiver operating characteristic curve results for each feature across each score on the Global Ratings of Change Scale.

Score on Global Ratings of Change Scale	Area Under the Curve (95% Confidence Interval)	Specificity	Sensitivity	Accuracy	Threshold
OVERALL SEVERITY					
1	0.77 (0.70-0.84)	1.00	0.51	0.58	7.5
2	0.75 (0.68-0.82)	0.95	0.53	0.61	7.5
3	0.73 (0.66-0.80)	0.84	0.58	0.69	10.5
4	0.72 (0.64-0.81)	0.88	0.57	0.76	14.5

5	0.71 (0.61-0.81)	0.79	0.65	0.75	12.5
6	0.78 (0.67-0.91)	0.81	0.75	0.80	16.5§
7	0.80 (0.62-0.97)	0.79	0.82	0.79	19.5
ROUGHNESS					
1	0.70 (0.61-0.78)	0.89	0.51	0.56	4.5
2	0.65 (0.57-0.73)	0.82	0.52	0.58	4.5
3	0.67 (0.59-0.75)	0.74	0.59	0.66	4.5
4	0.66 (0.57-0.75)	0.76	0.61	0.70	6.5
5	0.67 (0.57-0.78)	0.65	0.71	0.67	5.5
6	0.71 (0.57-0.85)	0.88	0.61	0.84	16.5§
7	0.64 (0.40-0.89)	0.86	0.55	0.84	17.5
BREATHINESS					
1	0.71 (0.63-0.80)	1.00	0.36	0.45	7.5
2	0.68 (0.60-0.76)	0.92	0.41	0.51	6.5
3	0.71 (0.64-0.78)	0.90	0.46	0.65	7.5
4	0.70 (0.62-0.79)	0.96	0.43	0.75	15.5§
5	0.71 (0.62-0.81)	0.83	0.57	0.76	9.5
6	0.79 (0.68-0.90)	0.82	0.71	0.80	10.5
7	0.79 (0.59-0.98)	0.92	0.73	0.90	28.5
STRAIN					
1	0.68 (0.59-0.77)	0.78	0.60	0.62	0.5
2	0.65 (0.57-0.74)	0.79	0.50	0.56	4.5
3	0.66 (0.58-0.74)	*0.68/0.72	*0.64/0.60	*0.66/0.65	*2.5/3.5

4	0.64 (0.56-0.73)	0.92	0.41	0.72	17
5	0.67 (0.56-0.78)	0.89	0.49	0.79	17
6	0.77 (0.65-0.89)	0.88	0.68	0.85	17
7	0.79 (0.62-0.96)	0.83	0.73	0.82	17

*ROC analyses resulted in two thresholds that were matched in the balance of sensitivity/specificity.
§ indicates MCID thresholds. Note: unable to establish a valid MCID for Strain.

DISCUSSION

The CAPE-V rating scale is a widely used tool for auditory-perceptual assessment of voice quality. It has the potential to be employed as a robust tool for determining changes in voice presentation following therapeutic, medical, or surgical intervention. However, no prior studies had investigated what degree of change in CAPE-V scores represents a true or clinically meaningful difference. The purpose of this study was to determine the sensitivity to change (i.e., MDC) and responsiveness (i.e., MCID) of the four voice *quality* features that comprise the CAPE-V index. To our knowledge, this study is the first attempt to empirically determine MDC or MCID in an auditory-perceptual voice scale, and one of the only MCID studies in voice research (Marks et al., 2021; Young et al., 2018).

MDC values for the voice quality subscales ranged from 12.06 mm (for Breathiness) to 18.74 mm (for Strain). These results indicate that scores less than these values cannot be confidently discriminated as a change, as they have a 95% chance of being within measurement error. From a subjective perspective, a clinician may feel as though a 12 to 18 mm change in any of the voice quality scales would be sufficient to constitute a change, but it is difficult to interpret subtle changes without an anchor. Thus, MCID values are crucial for interpreting changes. Even if a score is larger than the MDC, it might not be significant enough to indicate noticeable or

clinically important improvements. On the other hand, if a patient demonstrated less than a 12 to 18 mm change but reported amelioration or self-perceptually improved voice symptoms, clinicians should take this into consideration when interpreting changes.

In three out of four voice quality subscales in this study, the MCID was larger than the MDC, which is what we would expect based on the definitions of the concepts. In other words, theoretically, a clinically meaningful change cannot be smaller than measurement error (Riddle & Stratford, 2012). Since this study represents the first application of Jaeschke's global ratings of change scale (Jaeschke et al., 1989) to investigate changes in auditory-perceptual scales of voice features, we decided to take a data-driven approach to determine the MCID for each subscale, similar to that done by Stipancic et al. (2024). We determined the MCID based on which GROC level demonstrated optimized sensitivity and specificity *and* was higher than the MDC for each respective subscale. By doing this, we defined the magnitude of change required to be clinically meaningful while also being outside measurement error. For example, the *Overall Severity* subscale had an MDC value of 14.91 mm, suggesting anything larger than that is a true/real change. The MCID for this subscale was determined to be 16.5 mm, which was the threshold identified for the GROC level 6 (i.e., "a great deal better"). In other words, a change score on the Overall Severity subscale of 16.5 mm, which is greater than the threshold to define true change, is clinically meaningful based on a clinician's judgment of the patient's overall voice severity being "a great deal better." Strain was the one voice quality subscale for which we were unable to establish a valid MCID. This was not particularly surprising, given that Strain had the lowest reliability and greatest variability in both the current study and others (Zraick et al., 2011). Other MCID studies have found smaller MCID values compared to MDC values (Marks et al., 2021;

Stipancic et al., 2018). We recommend using the MDC as a threshold for real change in Strain until a valid MCID can be established (see Stratford & Riddle, 2012).

Overall Severity and Breathiness had the smallest MDCs and MCIDs compared to Roughness and Strain, which was expected based on literature that has described better reliability and less variability in the former two subscales compared to the latter two (Nagle, 2022). Since intra-rater reliability and standard deviation are part of the calculation of the SEM, the combination of lower reliability and larger variability translates to larger SEM values and subsequently larger MDC values. Standardized training procedures have been proposed for voice and speech auditory-perceptual analysis (Eadie & Baylor, 2006; Stipancic et al., 2023). Broadening access to training procedures for early clinicians and trainees should be emphasized, particularly for the Roughness and Strain scales. Having objective metrics to quantify real and clinically important auditory-perceptual changes has the potential to yield consistency across and within voice clinicians when assessing voice quality. Likewise, these findings could provide valuable insights for trainees and early clinicians to inform them of a patients' progress throughout voice therapy and/or discharge considerations. MDCs and MCIDs presented in this study would also provide context for patients on their progress through treatment, particularly if self-perception regarding voice changes is lacking.

Treatment efficacy studies often use the Overall Severity subscale as an outcome to document changes from treatment. Although most studies tend to report changes that are statistically significant, many of the average changes reported (i.e., values ranging from approximately 9 mm to 15 mm) are lower than the MDC and/or the MCID found in the current study (Braden & Thibeault, 2020; Fujiki et al., 2023; Reynolds et al., 2017). The distinction between statistically significant changes, those that are greater than measurement error, and

changes that are clinically meaningful is key when interpreting treatment efficacy studies. When therapy gains are larger than random measurement error, it is likely that a true change has occurred. Yet, when those gains do not exceed a clinically meaningful threshold, interpretation of those results should be tempered. For example, suppose an experimental treatment for patients with vocal fold nodules results in a decrease in Overall Severity by 15 mm, which is found to be statistically significant. This finding would be on the cusp of being considered a *true* change. However, based on the MCIDs established in the current study, this magnitude of change would not be considered clinically meaningful. Therefore, clinicians and researchers alike should use discretion when interpreting such findings.

Limitations and Future Directions

This study represents a first step into interpreting changes in voice quality, and caution should be exercised when using the MDCs and MCIDs across various voice diagnoses. Raters in this study were blinded to the patients' diagnoses (i.e., benign vocal fold lesions), so there could be the assumption that these scores could generalize across voice populations. However, findings from this study can only be confidently applied to patients with phonotrauma. Furthermore, any MDC or MCID can only be specific to the exact context in which the values are obtained, which includes the speakers, raters, listening task, etc. ICC calculations were made using an agreement approach, which is based on averages across raters. Since only three raters judged each set of recordings, it is possible that results can only be applied to the specific clinicians who participated in the study or generalized to ratings of three clinicians in clinical practice. Fortunately, inter-rater reliability scores (seen in Table 3) found good to excellent reliability for all of the voice quality dimensions except for Strain in one group of listeners. We feel that these cutoff values can be confidently used as a starting point for judging whether real and meaningful

change has occurred. To broaden the context in which MDCs and MCIDs for the CAPE-V can be applied, future work should investigate MDC and MCID values across various voice populations outside of phonotraumatic voice disorders, with a larger cohort of raters, and using a single-rater reliability calculation.

Another factor to consider is that speakers in this cohort were mostly categorized in the mild to moderate severity range. As our speaker cohort did not include an even amount of mild, moderate, or severe ratings, this uneven distribution could have influenced the current results. Patients with phonotrauma tend to score in the mild to moderate ranges on the CAPE-V (Behrman et al., 2004; Toles et al., 2021), even in "moderate-to-severe" manifestations of phonotrauma (Van Stan et al., 2023). The score distribution of the sample in the current study is in line with those previous reports. Despite the skewed distribution, there is reasonable variability within the perceptual features, with pre-treatment standard deviations ranging from 15 mm to 25 mm (i.e., encompassing one-third to one-half of the scale across perceptual features). Considering the distributional characteristics of the sample, the MDCs and MCIDs established by this work are primarily applicable to more mild-to-moderate dysphonia presentations.

It is also possible that the higher the severity rating, the higher the MDC or MCID necessary to quantify a clinically relevant change. This phenomenon is evident in related work, which found that severity of the disorder greatly affected estimates of MDCs and MCIDs in speech intelligibility (Stipancic & Tjaden, 2022; Stipancic et al., 2018). Conversely, since the sample was skewed toward the mild-to-moderately dysphonic range and CAPE-V scores are generally more reliable at the extreme ends of the scale, it is also possible that the MDCs/MCIDs found in the current study are larger than they would be if we had included more typical and severely dysphonic speakers (i.e., those who would be rated the extreme ends of the scale). As

such, future work should employ a cohort that involves speakers with voice qualities more evenly distributed across the severity continuum and investigate differences in MDCs and MCIDs between severity groups containing relatively similar numbers of speakers.

Audio recordings included for rating in this study were obtained during their visits with the laryngologist. Recordings were routinely conducted with some amount of background noise as the stroboscopy equipment was running in the background, and signal gain was not able to be controlled. Therefore, many of the audio recordings from the original cohort of patients had to be removed before finalizing the patient sample for rating. Though all recordings were deemed appropriate for rating by two of the authors (J.C.S. and L.E.T.), it is possible that the presence of background noise or differences in sound level could have affected results. Future work should use samples that are obtained in a more controlled environment.

The current study had two potential deviations from typical auditory-perceptual evaluation using the recommended CAPE-V protocol. First, the official CAPE-V protocol does not include evaluation of the voice during recitation of the Rainbow Passage. We decided to use the Rainbow Passage because it was the only standardized speech passage consistently available in this retrospective database. However, we felt that it was reasonable to judge voice quality using this passage since previous work has reported that SLPs regularly include ratings from the Rainbow Passage in their clinical practice (Lodhavia & Kempster, 2024; Nagle et al., 2024). In fact, a revised CAPE-V is currently under review recommends the use of the first three sentences of the Rainbow Passage for auditory-perceptual rating (Kempster et al., 2024 preprint). Second, we used a VAS that had linearly distributed severity markers (i.e., mild, moderate, and severe textual markers placed at equal increments along the scale). Though the original CAPE-V protocol developed from the consensus meeting in 2002 uses nonlinearly distributed markers, we

opted to use linearly placed markers found in the peer-reviewed version of the scale published in Kempster et al. (2009). Previous work has found that the presence and location of the markers can potentially affect ratings (Nagle et al., 2014). It is possible that we could have found differences in ratings, and thus MDC/MCID estimates, if we had used the version with nonlinearly distributed markers or a version with no severity markers (Walden & Rau, 2022; Kempster et al., 2024 preprint).

Patients who identified no changes through treatment or a worsening of voice quality were not included in the current study. Though this decision was made to be able to determine MDCs and MCIDs for *improvements* in voice quality, it precludes the ability to apply these thresholds to worsening voice quality. There are certainly times when treatment is not effective or conditions worsen, leading to voice quality deterioration. Further investigation into identifying what constitutes real and clinically meaningful changes in worsening voice quality is warranted.

The current study used an anchor-based approach to determine changes from the clinician's perspective. We felt that this was the best approach to use as a first step when determining changes on a clinician-rated scale. However, our study does not address the patient's perspective on their voice and voice problem. It is possible that some patients will perceive smaller changes as personally significant to them despite the clinician not recognizing a significant change. Conversely, some patients have less sensitive awareness (Smeltzer et al., 2023) and might not recognize an improvement until larger changes occur. Other perspectives to consider are care partners or naïve listeners, both of whom might provide additional insights into what is relevant in real life situations. Future work should investigate MDCs and MCIDs in the CAPE-V using patient-reported anchors such as self-report questionnaires or self-reported

ratings of dysphonia, ratings from untrained listeners, or ratings from individuals familiar with the patient's voice.

The question that we used for the GROC scale queried whether and to what degree the rater noticed a difference in the speaker's overall voice quality. One of the primary objectives in voice therapy and laryngeal surgery is to improve the patient's voice quality as a whole and not necessarily certain dimensions of voice quality (although, of course, in some cases specific dimensions might need to be targeted). Therefore, we believe that comparing improvements in each of the voice quality dimensions to a global measure of voice quality improvement is a valid way to determine changes. However, it is possible that this question was too broad to use as an anchor for each of the voice quality subscales and is possibly one explanation for the difficulty identifying a valid MCID for Strain. For example, consider a patient with moderate Overall Severity but without any perceptible Strain pre-treatment, yet was deemed to have a large improvement in overall voice quality following treatment. In this case, there would not be any change in Strain ratings despite the anchor measurement representing a large change, leading to no or small change in Strain being labeled as a large change. This problem could be addressed using separate GROC scales for each voice feature, which might yield different, and perhaps more accurate, MCIDs. For instance, future work could identify the MCID of the Strain subscale by specifically asking raters to discriminate the amount of strain in the voice quality that is present across voice samples using the GROC scale. Additionally, providing an anchor scale with fewer response options could prove to yield greater sensitivity and specificity, and be easier for listeners to employ, as suggested by Stipancic et al., (2024).

CONCLUSION

This study is the first known attempt to establish MDCs or MCIDs within the context of pre- and post- voice treatment outcome measures when measuring auditory-perceptual changes of voice quality. MDCs across voice quality parameters were 14.9 mm for Overall Severity, 14.6 mm for Roughness, 12.1 mm for Breathiness, and 18.7 mm for Strain. MCIDs were 16.5 mm for Overall Severity, 16.5 mm for Roughness, and 15.5 mm for Breathiness. A valid MCID for Strain was not obtained. These findings served to answer what degree of change on the voice quality dimensions of the CAPE-V scale constitutes real (i.e., the MDC) and clinically important changes (i.e., the MCID) for individuals undergoing treatment for phonotraumatic vocal fold lesions. Moreover, these findings provide a benchmark which clinicians can use to determine whether a patient has made objective progress in improving their voice quality throughout the course of voice therapy or following medical or surgical intervention. Though these findings provide objective information about meaningful changes in voice quality from the clinician perspective, clinicians must also consider the patient's perception of their symptoms when assessing progress and discharge from voice therapy.

ACKNOWLEDGMENTS

The authors would like to thank the speech-language pathologists who served as expert raters for this study. Special thanks to Drs. Lesley Childs and Ted Mau for providing access to their phonotrauma patient database.

FUNDING STATEMENT: This work was supported by NIH/NIDCD grant K23DC020758 (PI: L. Toles) and a New Investigators Research Grant from the American Speech-Language Hearing

650 Foundation (PI: L. Toles). The content is solely the responsibility of the authors and does not
651 necessarily represent the official views of the National Institute on Deafness and Other
652 Communication Disorders, the National Institutes of Health, or the ASHFoundation.

653

654 **DATA AVAILABILITY STATEMENT**

655 The datasets generated and/or analyzed during the current study are available from the
656 corresponding author upon reasonable request.

657

658

659 **REFERENCES**

660 Amit, M., Abergel, A., Fliss, D. M., & Gil, Z. (2012). The clinical importance of quality-of-life
661 scores in patients with skull base tumors: a meta-analysis and review of the literature.
662 *Current oncology reports*, *14*, 175-181.

663 Beckerman, H., Roebroeck, M., Lankhorst, G., Becher, J., Bezemer, P. D., & Verbeek, A.
664 (2001). Smallest real difference, a link between reproducibility and responsiveness.
665 *Quality of Life Research*, *10*, 571-578.

666 Behrman, A., Sulica, L., & He, T. (2004). Factors predicting patient perception of dysphonia
667 caused by benign vocal fold lesions. *The Laryngoscope*, *114*(10), 1693-1700.

668 Beninato, M., & Portney, L. G. (2011). Applying concepts of responsiveness to patient
669 management in neurologic physical therapy. *Journal of Neurologic Physical Therapy*,
670 *35*(2), 75-81.

671 Braden, M., & Thibeault, S. L. (2020). Outcomes of voice therapy in children with benign vocal
672 fold lesions. *International Journal of Pediatric Otorhinolaryngology*, *136*, 110121.

673 Carding, P. N., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes:
674 state of the science review. *The Journal of Laryngology & Otology*, *123*(8), 823-829.
675 Childs, L. F., D'Oto, A., Harris, A., Rao, A., & Mau, T. (2022). Voice Therapy Expectations for
676 Injured Singers. *Journal of Voice*.
677 Childs, L. F., Rao, A., & Mau, T. (2022). Profile of Injured Singers: Expectations and Insights.
678 *The Laryngoscope*. https://doi.org/10.1002/lary.30015
679 de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M.
680 (2006). Minimal changes in health status questionnaires: distinction between minimally
681 detectable change and minimally important change. *Health and quality of life outcomes*,
682 *4*, 1-5.
683 Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners'
684 judgments of dysphonic voice. *Journal of Voice*, *20*(4), 527-544.
685 http://linkinghub.elsevier.com/retrieve/pii/S0892199705000998?showall=true
686 Eadie, T. L., & Kapsner-Smith, M. (2011). The effect of listener experience and anchors on
687 judgments of dysphonia.
688 Fujiki, R. B., Braden, M., & Thibeault, S. L. (2023). Voice Therapy Improves Acoustic and
689 Auditory-Perceptual Outcomes in Children. *The Laryngoscope*, *133*(4), 977-983.
690 Fulk, G. D., & Echternach, J. L. (2008). Test-retest reliability and minimal detectable change of
691 gait speed in individuals undergoing rehabilitation after stroke. *Journal of Neurologic*
692 *Physical Therapy*, *32*(1), 8-13.
693 Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests.
694 *Logopedics Phoniatrics Vocology*, *28*(3), 109-116.

Haley, S. M., & Fragala-Pinkham, M. A. (2006). Interpreting change scores of tests and measures used in physical therapy. *Physical therapy*, *86*(5), 735-743.

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., & Kirby, J. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of biomedical informatics*, *95*, 103208.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*, *42*(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

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.

Hillman, R. E., Stepp, C. E., Van Stan, J. H., Zañartu, M., & Mehta, D. D. (2020). An updated theoretical framework for vocal hyperfunction. *American journal of speech-language pathology*, *29*(4), 2254-2260.

Holmberg, E. B., Hillman, R. E., Hammarberg, B., Södersten, M., & Doyle, P. (2001). Efficacy of a behaviorally based voice therapy protocol for vocal nodules. *Journal of Voice*, *15*(3), 395-412. <Go to ISI>://000171055000009

Horváth, K., Aschermann, Z., Kovács, M., Makkos, A., Harmat, M., Janszky, J., Komoly, S., Karádi, K., & Kovacs, N. (2017). Changes in quality of life in Parkinson's disease: how large must they be to be relevant? *Neuroepidemiology*, *48*(1-2), 1-8.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Controlled clinical trials*, *10*(4), 407-415.

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. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18930908

Kempster, G. B., Nagle, K. F., & Solomon, N. P. (2024, November 12). Development and Rationale for the Consensus Auditory-Perceptual Evaluation of Voice – Revised (CAPE-Vr). https://doi.org/10.31234/osf.io/e84tn

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.

Koopman, J. E., van Kooij, Y. E., Selles, R. W., Slijper, H. P., Smit, J. M., van Nieuwenhoven, C. A., Wouters, R. M., & Group, H.-W. S. (2023). Determining the Minimally Important Change of the Michigan Hand outcomes Questionnaire in patients undergoing trigger finger release. *Journal of Hand Therapy*, *36*(1), 139-147.

Kreiman, J., & Gerratt, B. R. (2010). Perceptual assessment of voice quality: Past, present, and future. *Perspectives on Voice and Voice Disorders*, *20*(2), 62-67. https://doi.org/10.1044/vvd20.2.62

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. <Go to ISI>://A1993KM22300003

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, *35*(3), 512-520.

Lodhavia, A., & Kempster, G. B. (2024). Fidelity to the Consensus Auditory-Perceptual Analysis of Voice (CAPE-V): A Pilot Study. *Journal of Voice*.

Marks, K. L., Verdi, A., Toles, L. E., Stipancic, K. L., Ortiz, A. J., Hillman, R. E., & Mehta, D. D. (2021). Psychometric Analysis of an Ecological Vocal Effort Scale in Individuals With and Without Vocal Hyperfunction During Activities of Daily Living. *American journal of speech-language pathology*, *30*(6), 2589-2604.

Merfeld, D. M. (2011). Signal detection theory and vestibular thresholds: I. Basic theory and practical considerations. *Experimental brain research*, *210*(3), 389-405.

Nagle, K. F. (2016). Emerging scientist: Challenges to CAPE-V as a standard. *Perspectives of the ASHA Special Interest Groups*, *1*(3), 47-53. https://doi.org/10.1044/persp1.SIG3.47

Nagle, K. F. (2022). Clinical Use of the CAPE-V Scales: Agreement, Reliability and Notes on Voice Quality. *Journal of Voice*.

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.

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*.

Nemr, K., Simoes-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *Journal of Voice*, *26*(6), 812. e817-812. e822.

Park, Y., Anand, S., Gifford, S. M., Shrivastav, R., & Eddins, D. A. (2023). Development and validation of a single-variable comparison stimulus for matching strained voice quality using a psychoacoustic framework. *Journal of Speech, Language, and Hearing Research*, *66*(1), 16-29.

Patel, S., Shrivastav, R., & Eddins, D. A. (2012). Identifying a comparison for matching rough voice quality. *Journal of Speech, Language, and Hearing Research*, *55*(5), 1407-1422. https://doi.org/10.1044/1092-4388(2012/11-0160)

Poulton, E. C. (2023). *Bias in quantifying judgments*. Routledge.

Reynolds, V., Meldrum, S., Simmer, K., Vijayasekaran, S., & French, N. (2017). A randomized, controlled trial of behavioral voice therapy for dysphonia related to prematurity of birth. *Journal of Voice*, *31*(2), 247. e249-247. e217.

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-based clinical voice assessment: A systematic review. *American journal of speech-language pathology*, *22*(2), 212-226. https://doi.org/10.1044/1058-0360(2012/12-0014)

Roy, N., Merrill, R. M., Gray, S. D., & Smith, E. M. (2005). Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *Laryngoscope*, *115*(11), 1988-1995. <Go to ISI>://000233839600016

Salaffi, F., Stancati, A., Silvestri, C. A., Ciapetti, A., & Grassi, W. (2004). Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *European journal of pain*, *8*(4), 283-291.

Sauder, C. L., Kapsner-Smith, M. R., 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.

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, *48*(2), 323-335. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15989395

Smeltzer, J. C., Chiou, S. H., & Shembel, A. C. (2023). Interoception, voice symptom reporting, and voice disorders. *Journal of Voice*.

Snook, S. H. (1999). Future directions of psychophysical studies. *Scandinavian Journal of Work, Environment & Health*, 13-18.

Stipancic, K. L., Brenk, F., Qiu, M., & Tjaden, K. (2024). Progress Toward Estimating the Minimal Clinically Important Difference of Intelligibility: A Crowdsourced Perceptual Experiment. *Journal of Speech, Language, and Hearing Research*, 1-15.

Stipancic, K. L., Golzy, M., Zhao, Y., Pinkerton, L., Rohl, A., & Kuruvilla-Dugdale, M. (2023). Improving perceptual speech ratings: The effects of auditory training on judgments of dysarthric speech. *Journal of Speech, Language, and Hearing Research*, *66*(11), 4236-4258.

804 Stipancic, K. L., & Tjaden, K. (2022). Minimally detectable change of speech intelligibility in
805 speakers with multiple sclerosis and Parkinson's disease. *Journal of Speech, Language,*
806 *and Hearing Research*, *65*(5), 1858-1866.
807 Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2018). Minimally detectable change
808 and minimal clinically important difference of a decline in sentence intelligibility and
809 speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech,*
810 *Language, and Hearing Research*, *61*(11), 2757-2771.
811 Stratford, P. W., & Riddle, D. L. (2012). When minimal detectable change exceeds a diagnostic
812 test–based threshold change value for an outcome measure: Resolving the conflict.
813 *Physical therapy*, *92*(10), 1338-1347.
814 Toles, L. E., Ortiz, A. J., Marks, K. L., Burns, J. A., Hron, T., Van Stan, J. H., ... & Hillman, R.
815 E. (2021). Differences between female singers with phonotrauma and vocally healthy
816 matched controls in singing and speaking voice use during 1 week of ambulatory
817 monitoring. *American Journal of Speech-Language Pathology*, *30*(1), 199-209.
818 Van Stan, J. H., Burns, J., Hron, T., Zeitels, S., Panuganti, B. A., Purnell, P. R., ... &
819 Ghasemzadeh, H. (2023). Detecting mild phonotrauma in daily life. *The*
820 *Laryngoscope*, *133*(11), 3094-3099.
821 Walden, P. R., & Rau, S. (2022). Individual voice dimensions' prediction of overall dysphonia
822 severity on two auditory-perceptual scales. *Journal of Speech, Language, and Hearing*
823 *Research*, *65*(8), 2759-2777.
824 Wright, C. J., Linens, S. W., & Cain, M. S. (2017). Establishing the minimal clinical important
825 difference and minimal detectable change for the Cumberland Ankle Instability Tool.
826 *Archives of physical medicine and rehabilitation*, *98*(9), 1806-1811.

Young, V. N., Jeong, K., Rothenberger, S. D., Gillespie, A. I., Smith, L. J., Gartner-Schmidt, J. L., & Rosen, C. A. (2018). Minimal clinically important difference of voice handicap index-10 in vocal fold paralysis. *The Laryngoscope*, *128*(6), 1419-1424.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, 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)