Clinical Use of the CAPE-V Scales: Agreement, Reliability and Notes on Voice Quality

Kathleen F. Nagle, *Nutley, New Jersey*

Summary: Objectives. The CAPE-V is a widely used protocol developed to help standardize the evaluation of voice. Variability of voice quality ratings has prevented development of training protocols that might themselves improve interrater agreement among new clinicians. As part of a larger mixed methods project, this study examines agreement and reliability for experienced clinicians using the CAPE-V scales.

Study Design. Observational.

Methods. Experienced voice clinicians (N=20) provided ratings of recordings from 12 speakers representing a range of overall voice quality. Participants were instructed to rate the voices as they normally would, using the CAPE-V scales. Descriptive data were recorded and two levels of agreement were calculated. Single rater reliability was calculated using a 2-way random model of absolute agreement for intraclass correlations (ICC [2,1]).

Results. Participants use of the CAPE-V scales varied considerably, although most rated overall severity, breathiness, roughness and strain. Data from one participant did not meet *a priori* agreement criteria. Because outcomes were significantly different without their data, agreement and reliability were analyzed based on the reduced data set from 19 participants. Interrater agreement and reliability were comparable to previous research; the mean range of ratings was at least 47mm for all dimensions of voice quality.

Conclusions. Results indicated differential use of the components of the CAPE-V form and scales in evaluating voice quality and severity of dysphonia, including categorical variability among ratings of all of the primary CAPE-V dimensions of voice quality that may complicate the clinical description of a voice as mildly, moderately or severely dysphonic.

Key Words: Evaluation−CAPE-V−Auditory-perceptual−Agreement−Reliability.

INTRODUCTION

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) is a widely-used clinical instrument that includes a protocol for obtaining voice samples and a method for clinicians to rate auditory-perceptual dimensions of voice quality. It was originally developed by the ASHA Special Interest Division 3 (now SIG3: Voice and Upper Airway Disorders) to standardize the terminology used to describe voice quality and to propose a psychometrically valid method of perceptual voice evaluation.[1] Important benchmarks of the utility of a rating system include rater reliability, sensitivity to differences, and agreement. Initial studies introducing the CAPE-V reported sufficient interrater reliability among clinicians to evaluate voice quality differences;[2,3] however, measures of interrater agreement have revealed considerable variability among voice clinicians using the CAPE-V scales.[4,5]

The issue of interrater variability hampers the use of the CAPE-V as a standard clinical tool.[6] It prevents comparison of ratings provided by different raters, but more importantly, it complicates identification of exemplar samples that may be used to train clinicians or to serve as an external standard to be shared by all voice clinicians. The problem is circular: Identifying voice samples to be used as auditory anchors requires consensus on terminology and use of the CAPE-V rating scales, but users of the CAPE-V may require training to reach or to provide consensus on what those samples might be. Because auditory-perceptual characteristics are a primary factor in characterizing disordered voice,[7] even incremental improvements in interrater agreement on these characteristics are desirable as indicators of baseline impressions and change over time. Such improvements may be achievable through the use of auditory anchors with corresponding voice quality judgments from experienced voice clinicians using the CAPE-V.

Before we can identify or develop valid auditory anchor samples for the CAPE-V, it is critical to investigate how experienced voice clinicians actually use the CAPE-V scales to rate voice quality. Selection of a set of auditory-perceptual anchors will require actual consensus on ratings of overall severity, breathiness, roughness, and strain for voices with a variety of types of dysphonia. Ideally, experienced voice clinicians would evaluate a set of voices and agree on the "degree of perceived deviance from normal for each parameter. . .."[1] However, there is evidence that even experienced voice clinicians may not use the CAPE-V rating scales as directed, based on wide ranges of absolute CAPE-V ratings reported in research by authors such as Walden (2020). In a pilot survey of 17 members of ASHA's Special Interest Group 3, Voice and Airway Disorders, only 41.3% of respondents reported administering all of the components

Accepted for publication November 10, 2022.
I have no conflicts of interest to disclose.
This work was funded by a 2018 American-Speech-Language-Hearing Foundation New Investigators Research Grant.
From the Department of Speech-Language Pathology, School of Health & Medical Science, Seton Hall University, Nutley, New Jersey.
Address correspondence and reprint requests to Kathleen F. Nagle, Department of Speech-Language Pathology, School of Health & Medical Science, Seton Hall University, 123 Metro Boulevard, Room 0440, Nutley, NJ, 07110 E-mail: naglekat@shu.edu
Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2022.11.014

of the CAPE-V, and only three of them reported that they always follow the protocol for marking the scales.[8]

This paper reports the results of the first part of a mixed methods project examining how experienced voice clinicians use the CAPE-V in practice. First, challenges to gaining consensus on absolute ratings of multidimensional auditory-perceptual stimuli for the CAPE-V scales are reviewed. This review is followed by an observational study of how experienced voice clinicians used the CAPE-V form and scales to evaluate a set of voice samples, with a discussion of how they rated them and how this information might be used to establish auditory anchors for use by all clinicians.

Challenges to consensus

Identifying external standards

All listeners have an internal standard of voice quality against which they judge any voice sample.[9] These standards are unstable and can result in very different absolute quality ratings for the same voice sample over time. Training with exemplars ranging in degree for four primary dimensions of voice quality (overall severity, breathiness, roughness, strain) is one way to increase rater reliability. Currently, however, no single "score" can be truly assigned to a voice sample, because of the multidimensional nature of voice; the systematic differences with which even experts approach rating voice; and the variety of scaling methods used to obtain a score. With no external standards beyond the textual markers of "mild," "moderate" and "severe," the CAPE-V protocol itself provides no means of increasing agreement among raters, and even these text anchors may force auditory-perceptual judgments into more categorical distinctions than might otherwise be noted by a given rater. The wide variability reported among even "expert" raters on any one voice sample has so far precluded identification of valid auditory anchor samples.[4,10]

The ability to assign a reliable quantitative measure of clinically relevant voice parameters to a patient's voice is critical to providing evidence-based care. Auditory-perceptual evaluations are in some cases critical for differential diagnoses between voice disorders (e.g., muscle tension dysphonia vs. spasmodic dysphonia), but most often serve as baseline data that may be compared after behavioral or medical intervention to measure treatment effects. Interrater differences in ratings may actually exceed the differences needed to document treatment effects or to classify dysphonic voices. Discrepancies would likely be minimized if users were trained specifically in the use of the CAPE-V, particularly if auditory anchors were contemporaneously available during clinical assessment of voice quality. Even experienced listeners have demonstrated significantly better interrater agreement for ratings of strain and overall severity when given access to auditory anchor samples.[4]

One of the initial goals of the development of the CAPE-V, therefore, was to create exemplars for use as auditory anchors and for training purposes.[1] To some extent, the University of Wisconsin Voice Disorders: Simulations (UWVDS; https://slpsims.csd.wisc.edu/) and the Perceptual Voice Qualities Database (PVQD)[11] have accomplished this goal. The UWVDS provides recordings of 45 speakers with a range of voice quality across CAPE-V parameters, who produced all of the tasks from the CAPE-V protocol (i.e., sustained vowels /a, i/; six sentences designed to a variety of vocal behaviors; and a brief sample of spontaneous speech). Users can rate and compare their ratings of each voice to those of an expert voice clinician. The PVQD consists of productions of sustained vowels and CAPE-V sentences from 296 speakers (89 with no voice complaint; 187 with either a voice complaint or a confirmed diagnosis of dysphonia). Each speaker's voice quality was rated by three or four listeners with at least 2 years' experience working with voice disorders on at least a monthly basis, and 2) familiarity with both the CAPE-V and GRBAS[12] rating scales, also used on at least a monthly basis.

The recordings of the PQVD and UWVDS present a golden opportunity to identify a range of truly representative auditory anchors. These resources provide freely available samples of healthy and dysphonic voice accompanied by expert ratings and are a valuable asset to clinicians wishing to refine their evaluation of voice quality. However, there are some limitations to treating either set of recordings as exemplars. Recordings were made in a clinical environment, and some contain residual background or recording noise. Very few pediatric voices are included in either set. UWVDS ratings were made by a single clinician, and may reflect that clinician's individual bias. Seeing an absolute value attached to a given voice may suggest to inexperienced clinicians that a single "correct" rating exists for any voice. PVQD ratings were obtained using traditional unmarked visual analog (VA) scales (i.e., not the hybrid VA scales with textual markers used on the CAPE-V form). Although this was done to facilitate comparison with GRBAS ratings, it affects the generalizability of the results for any individual voice recording, given potential effects of textual markers on auditory-perceptual ratings of voice quality.[13] Interrater reliability among three or four raters was calculated using a 1-way ANOVA, which provides a measure of the degree of consistency between (or within) those raters, and reported as ICC (1, k) = .86. This was necessary because different sets of raters were used for each set of stimuli. A review of the Ratings Spreadsheets provided as part of the PVQD[14] also indicates a wide range of absolute ratings. For example, ratings of overall severity varied by an average of 30.41mm, with some differing by more than 60mm on the 100mm VA scale. The main drawback to using the PQVD or UWVDS as exemplars, however, is the small number of ratings obtained for each stimulus (i.e. less than five). When interrater variability is relatively high, it may be best to establish an average score based on a larger number of ratings (possibly including multiple ratings by the same raters), and provide the range of those ratings to the user.[15]

Another way to address the need for "standard" auditory anchors might be to create them. Efforts to model the percepts of breathiness,[16] roughness[17] and strain[18] using single

variable matching tasks have resulted in the recent development of psychophysical scales that show great promise as clinical tools.[19] Synthetic stimuli developed from this research would provide auditory anchors with known physical and psychophysical properties, such that the multidimensional factors affecting voice quality dimensions could be controlled or varied to mimic, for example, a voice that is severely rough, moderately breathy and mildly strained, at least in theory.[10] This research is still in development, however, and currently relies on relatively simple sustained vowel stimuli.

In summary, gaining consensus on ratings of voice quality requires exemplars on which those ratings can be based. Obtaining a large number of ratings and reporting both averages and ranges of ratings for stimuli would provide functional criteria for training and clinical reference.

Design of the CAPE-V

Perceptual rating scales are ubiquitous in the assessment of speech and voice, but their appropriate use depends somewhat on the percepts being measured. In brief, rating scales are used to transform an auditory stimulus to a numerical representation of human perception of the stimulus. They may be described in terms of what the rater is asked to do; for example, Stevens[20] divided perceptual scales into confusion, partition and magnitude types. With a confusion scale, for example, raters are asked to identify just noticeable differences (JNDs) between unidimensional stimuli. Partition tasks require listeners to place a stimulus on a categorical, ordinal or interval scale. All observations within the same category are considered equal; no judgment of the quantity or severity of the dimension is made. Ordinal scales (e.g., "Mild, Moderate, Severe") presume that data can be ranked by quantity or severity, but provide no indication of how "far apart" the categories might be. Interval scales have defined units of measure; the distance between points on the scale has meaning, and can be assumed to be equivalent along the length of the scale. Equal appearing interval (EAI) scales (i.e., X-point scales, which cannot be further partitioned) are an example.

Because of the nonlinearity of some auditory-perceptual dimensions of voice, certain types of data should not be scaled with partition scales.[21–23] For these dimensions, listener discrimination varies from one end of the scale to the other, meaning that raters cannot psychometrically divide the scale into equal intervals. In such cases, direct magnitude estimation (DME) or traditional visual analog (VA) scales should be used.[24] A traditional VA is an undifferentiated line (i.e., with no partitions), and listeners assign ratings in proportion to their perceived magnitude. Because of the potential differences in discrimination across magnitude scales, comparison of ratings among listeners is not as straightforward as it is with partition scales. Both partition and magnitude tasks are commonly used in measuring speech and voice stimuli.[25,26] For a full treatment on considerations of perceptual scales, the reader is directed to Shrivastav, Sapienza and Nandur.[15]

The CAPE-V rating scales were developed based on a pragmatic approach to improving consistency of voice evaluation among clinicians. The hybrid scales were designed to optimize reliability within and between clinicians in a protocol that could be efficiently executed.[1] To that end, each 100mm CAPE-V scale elicits interval data measured in millimeters, and text markers beneath each line indicate roughly specified increasing (ordinal) categories of severity (Figure 1A). Text markers were designed to appear with "MI/MO/SE" centered at about 10mm, 50mm and 90mm, respectively, and there are no labels at the endpoints. Thus, the CAPE-V scales contain aspects of ordinal, interval and ratio scales. Users are asked to partition the scale, but also to indicate the magnitude of the perceptual dimension being rated.

The CAPE-V scales represent an attempt to apply psychometric standards to measurement of multidimensional stimuli for which the salience of specific attributes varies among raters. These scales were designed to be useable across a range of vocal pathologies by clinicians with varying degrees of experience,[1] and they signify a substantial improvement over ordinal auditory-perceptual voice quality measurement tools such as the GRBAS[12] and Buffalo Voice Profile.[27] It can be challenging to determine how to interpret CAPE-V ratings, however, at least in part because it may not be appropriate psychometrically to measure breathiness, roughness, strain or overall severity of dysphonic voice using partition scales.

In summary, the undefined nature of the ordinal text markers (and their placement) complicates the interpretation of CAPE-V ratings from one clinician to the next. Exploring how voice clinicians use the CAPE-V scales in practice is critical to developing or identifying exemplars for training and clinical reference.

Reliability and agreement

In addition to variables related to the task of rating, auditory-perceptual ratings are necessarily influenced by factors beyond the characteristics of the voice signal. Listeners bring their own states and traits to the task of perceiving a stimulus and assigning a rating to it, such as hearing status, familiarity with the speaker, and experience with dysphonic voices.[28,29] Listeners in an auditory-perceptual task are the tools of measurement; to provide valid outcomes, their judgments should be calibrated as closely as possible within the constraints of these individual listener differences. To interpret auditory-perceptual data, it is critical to report reliability and agreement within and among raters.

As described by Kreiman and colleagues, "Ratings are *reliable* when the relationship of one rated voice to another is constant (i.e., when voice ratings are parallel or correlated), although the absolute rating may differ from listener to listener (p.36)."[28] High interrater reliability (i.e., $r > .95$) has been reported for both experienced and inexperienced

CAPE-V Form

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Name:_____ **Date:**_____

The following parameters of voice quality will be rated upon completion of the following tasks:
1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent I = Intermittent
MI = Mildly Deviant MO = Moderately Deviant SE = Severely Deviant
Although the PDF scale is accurate, printer configurations vary. Verify that your paper copy has accurate 100-mm lines before reproducing this form.

Overall Severity _____ C I /100
MI MO SE

Roughness _____ C I /100
MI MO SE

Breathiness _____ C I /100
MI MO SE

Strain _____ C I /100
MI MO SE

Pitch (A) (Indicate the nature of the abnormality): _____
_____ C I /100
MI MO SE

(E)

Loudness (B) (Indicate the nature of the abnormality): _____
_____ C I /100
MI MO SE

(C) _____ C I /100
MI MO SE

_____ C I /100
MI MO SE

COMMENTS ABOUT RESONANCE: NORMAL OTHER (Provide description):_____

_____

ADDITIONAL FEATURES (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, or other relevant terms):

(D) Clinician:_____

Note. This form may be photocopied for clinical purposes.

FIGURE 1. CAPE-V form (initial version) showing scales and space for: **(A)** indicating pitch abnormality; **(B)** indicating loudness abnormality; **(C)** other factors affecting voice; **(D)** additional comments; and **(E)** indicating consistency and intermittency of a given dimension of voice. Figure adapted with permission from ASHA (2009); copyright 2009 ASHA.

raters judging overall severity of dysphonia;[30,31] however, these figures are based on Cronbach's alpha or average-measures intraclass correlations (ICC), which measure the association between each rating and the mean of every other rating for a given dimension. Kreiman et al.[28] refer to this as the reliability of the average rating; in Shrout & Fleiss terminology, this is ICC model (2, *k*).

Single-measures ICCs reflect the average agreement between one rater and another, or "single rater reliability", and reported as ICC (2,1) using the terminology of Shrout

and Fleiss.[32] Reliability of voice quality judgments from more than two raters on the CAPE-V is reported as single-measures ICC when examining the similarity of individual ratings within a group of listeners.[28] For example, Helou et al.[5] reported significantly reduced interrater reliability for inexperienced listeners (ICC[2,1] = .53) compared to experienced listeners (ICC [2,1]= .73) rating overall severity using the CAPE-V for samples of post thyroidectomy speech. In the validation study of the CAPE-V, single-measures ICCs between .56 and .76 were reported for 21 experienced clinicians, indicating moderate single rater reliability.[3,33]

Interrater agreement is generally reported as the probability of two ratings occurring within a given range on a given scale. Agreement criteria vary depending on the parameter being measured and the type of scale used. Ordinal scale data from the GRBAS, for example, can be analyzed in terms of Cohen's kappa, which can be weighted to account for chance agreement.[34] Equal-appearing interval (EAI) scales give raters a limited number of options, typically 5-9 points, although there could be more or fewer. The non-continuous nature of EAI scales makes their agreement criteria relatively straightforward; agreement is usually reported as the proportion of ratings falling exactly on or "within one scale value" of a given rating. Identification of agreement criteria for VA scales is somewhat arbitrary, however, and researchers have addressed this by evaluating agreement for VA scale data using criteria derived from an EAI scale.[25,28] For example, when evaluating interrater agreement and reliability among voice clinicians, Kreiman and colleagues (1993) compared EAI and VA scales, choosing a 7-point EAI scale because of its reported prevalence in clinical practice. They divided the 100mm VA scale into seven intervals of 14.3mm for comparison to the 7-point EAI scale.[28] The probability of "exact" agreement can be calculated for ratings within +/- 7.14mm (i.e., 14.3mm) on the VA scale, which represents one interval on a 7-point EAI scale. Likewise, ratings within +/- 21.5mm on the VA scale (i.e., three intervals including two points on a 7-point scale) may be considered to agree "within one scale value" in either direction of a given rating, in the terminology of Kreiman and colleagues.

These agreement criteria can be interpreted at each level in terms of chance. Given a uniform probability distribution, the probability of ratings falling within +/- 7.14mm of one another, which represents a potential range of 14.28mm of the scale, is 14% due to chance alone; the probability of chance agreement within +/- 21.5mm (i.e., potentially 43mm, or 43% of the scale) is 39%. For example, sentence-level productions presented to 10 experienced listeners showed marginally higher probabilities of "exact" agreement for vocal effort/strain (; 38%) and overall severity (34%) and "within one scale value" for strain (68%) and overall severity (75%) based on the same criteria.[4] Kreiman and colleagues[28] reported interrater agreement from 10 experienced listeners judging 22 pairs of pathological voice productions of a sustained vowel. The probability of exact agreement for these listeners was 28.4% (*SD* 8.05%), with a range of 10.0% to 46.7%; within one scale value, the probability of agreement for these voices was 68.8% (*SD* 13.01%), with a range of 33.3% to 93.3%. Mean listener agreement exceeded chance, arguably a low bar for expert listeners; however, at least one rater failed to do so at each level of agreement. A listener whose judgments differ that much from the group may be displaying systematic biases that require examination. Although arbitrary, an agreement criterion of exceeding chance probability seems justified for inclusion of individual rater data when the goal is to assign an absolute value or range of values to a given set of stimuli.

In summary, as a commonly used outcome measure developed by ASHA, the CAPE-V would benefit from the addition of evidence-based auditory standards, derived from a group of experienced voice clinicians, with maximal agreement on judgments of voice quality. Any such evidence would require fidelity to the CAPE-V protocol, or at least evidence of how implementation of the protocol varies among clinicians.

As part of a larger project investigating how experienced voice clinicians use the CAPE-V, this study aims to describe how voice clinicians use the CAPE-V scales in their daily practice, including the reliability and agreement of their voice quality ratings of a small set of samples, as a starting point for establishing a set of auditory anchors.

METHOD

This study reflects observations of how experienced voice clinicians use the CAPE-V form in daily practice. The context of data collection is important because certain factors that would normally be controlled for research purposes (e.g., listening conditions; instructions to raters; number of stimulus presentations) were allowed to vary to accommodate participants' preferences. Participants were asked to rate a small set of stimuli from the CAPE-V protocol using the CAPE-V scales (Kempster et al., 2009) as they normally would in their work setting. This study reports the following observations from these sessions:

1. Use of the CAPE-V form and scales: Whether and how clinicians marked ratings of voice quality (i.e., overall severity, breathiness, roughness, strain), loudness and pitch; additional features of voice quality; and consistency of these parameters;
2. Reliability among these participants when using the CAPE-V scales *as they normally would* to rate primary dimensions of voice quality (i.e., single rater reliability); and
3. Ranges and probability of interrater agreement on CAPE-V scale ratings for the primary dimensions of voice quality at two levels (i.e., exact agreement and the probability of agreement within one scale value).

The ratings obtained as part of this project represent a portion of the data obtained during 2-hour interviews

conducted in person at the participants' office or home, and which include qualitative observations about the use of the larger CAPE-V protocol and auditory-perceptual evaluation in general. Qualitative data will be reported separately.

Procedures were approved by the Seton Hall University Institutional Review Board and all participants were paid for their time.

Participants

Twenty voice clinicians (14 women, 6 men) were directly recruited from the author's professional network. To participate in the study, they had to have at least 3 years of current or recent experience evaluating voice quality using the CAPE-V, with a patient population of least 20% voice.[11,35] They ranged in age from 26-54 years (*M* 37, *SD* 8.4). All reported no hearing concerns. Clinicians were recruited from four broad geographical areas in the U.S. and represented different professional settings and backgrounds (Table 1). All had at least a master's level degree in speech-language pathology and were certified by the American Speech-Language-Hearing Association. Three had also obtained a PhD at the time of data collection. Additional information about their practice is shown in Table 1.

Stimuli

Stimuli were selected from a database of clinical voice evaluation recordings to represent a variety of voice impairments (i.e., dysphonias) and a range of dysphonia severity, based on expert clinician ratings of overall severity at the time of recording.[36] Details regarding preparation of stimuli and range of speaker diagnoses can be found in Table 1 of that paper (p.745).[36] The speaker group for the current study comprised two healthy speakers (1 man, 1 woman) between ages 25 and 32, and ten dysphonic speakers (5 men, 5 women) between ages 21 and 78. Because interrater agreement is known to decrease in the midrange of voice quality,[28,37] six of the speakers chosen for this study had been judged as moderately dysphonic; two had been judged as mild and two as severe (based on overall severity ratings[36]). Degree of overall severity was sex-matched, and is displayed in Table 2. Recordings had been collected with the CAPE-V protocol as part of a routine clinical voice evaluation, and included productions of 1) sustained vowel /a/ and 2) four of the CAPE-V sentences (i.e., "*How hard did he hit him," "We were away a year ago," "We eat eggs every Easter," "Peter will keep at the peak"*). Recordings were arranged pseudo-randomly by speaker to avoid presentation of more than two speakers with similar levels of overall severity in a row. Speakers are labeled in the order of presentation.

Task

Sessions occurred in a quiet room at the participants' home or office. Stimuli were presented on a laptop computer using Windows Media Player in a free field at a comfortable

TABLE 1.
Demographic Information for Experienced Participants

	Number of Participants	% of All Participants
Setting		
Academia	3	15%
Hospital/voice clinic	15	75%
Private practice	2	10%
Location		
Mid-Atlantic	5	25%
New England	6	30%
Pacific Northwest	5	25%
Southeast	4	20%
Voice experience (yrs.; *M* 8.15, *Med* 7.0, *SD* 4.98)		
3-5	9	45%
6-10	6	30%
11-20	5	25%
Voice evaluations per week (*M* 9.13, *Med* 5.5, *SD* 8.68)		
0-2	4	20%
3-5	6	30%
6-10	4	20%
11-20	4	20%
21-30	2	10%
Voice proportion of practice (*M* 83.5%, *Med* 95%, *SD* 25.96%)		
20%	2	10%
50-70%	3	15%
90% or greater	15	75%

TABLE 2.
Overall Severity of Dysphonia for Each Speaker, Based on Previous Ordinal Rating by Experienced Clinician; and Mean, Standard Deviation and Range of Ratings for Each Speaker, in mm for Reduced Data Set

Speaker	Sex	Overall Severity (N=1) *(original judgment)*	Overall Severity (n=19)			Breathiness (n=17)			Roughness (n=17)			Strain (n=17)		
			M	*(SD)*	Range	*M*	*(SD)*	Range	*M*	*(SD)*	Range	*M*	*(SD)*	Range
Sp1	Female	Mild	8.5	4.63	18	8.76	11.5	45	6	4.26	14	5.65	6.89	20
Sp2	Male	Moderate	41.89	16.1	67	36.24	17.14	74	28.35	17.34	61	24	16.02	59
Sp3	Male	Severe	68.95	13.53	43	54.29	19.57	65	52.59	21.07	76	61.94	19.13	70
Sp4	Female	Severe	55.11	15.85	51	57.24	17.23	58	24.24	17.77	50	24.76	17.33	66
Sp5	Male	Normal voice	14	12.9	60	1.41	3.04	10	13.18	13.86	60	9.59	11.34	45
Sp6	Female	Moderate	37.42	19.07	71	29.29	18.52	74	15.71	19.06	62	31.65	21.78	77
Sp7	Female	Normal voice	9.89	7.7	26	5.59	6.3	18	5.12	7.12	26	3.35	5.3	18
Sp8	Female	Moderate	79.68	12.36	49	65.24	16.38	52	50.82	25	81	69.65	19.07	78
Sp9	Male	Moderate	58.74	15.47	49	55.47	17.76	64	46.18	22.41	63	34.65	18.51	64
Sp10	Male	Mild	17.47	12.17	46	9.12	9.65	28	12.76	10.73	43	8.59	10.05	39
Sp11	Male	Moderate	57.84	15.97	54	12.41	16.02	55	55.76	18.42	62	25.29	22.07	70
Sp12	Female	Moderate	25.53	10.24	36	21.06	11.23	45	16.53	13.08	54	15.94	16.71	53
Mean					47.50			49.00			54.33			54.92
Median					49.00			53.50			60.50			61.50

volume set by each participant during the first speaker trial. All participants heard speakers in the same sequence, and all were permitted to hear as many repetitions as they requested. They were asked to rate the speaker the way they normally would in their clinical setting, given the limitations of doing so based on an audio recording rather than a live patient.

Although some clinics reportedly use computerized versions of the CAPE-V scales, the CAPE-V form was designed as a paper document. Participants used a paper form to evaluate these samples, as described by the authors of the CAPE-V[1] (Figure 1). Scales were verified as 100mm in length before use, in case of printing or photocopying adjustments to the original form.

Statistical analysis

Ratings were measured with a ruler and entered into a spreadsheet by the author and a research assistant, with an *a priori* criterion of 95% agreement on 10% of ratings (n=96) for inclusion in data analysis. Interrater agreement of these measurements exceeded 99%, with a mean difference of 0.13mm. Three participants provided both marking and numeric ratings for at least some of the speakers. When asked which method they used in clinic, they all indicated that the number should take precedence. Descriptive statistics were calculated per speaker for the four main dimensions of voice quality (i.e., overall severity, breathiness, roughness and strain). Markings of severity and consistency of dimensions of voice quality, loudness and pitch, and of other factors affecting voice that could be added to the CAPE-V form, are reported as proportions and frequency counts.

Interrater reliability was calculated with intraclass correlations (ICC), using a 2-way mixed model with raters as a random factor (ICC [2,1]) for "single rater reliability,"[32] based on the number of clinicians who provided ratings for each dimension. Intra-rater reliability was not calculated for this study. All statistics were calculated using SPSS version 28.0.

As described above, interrater agreement was determined in terms of the probability of exact agreement (P_{exact}; within +/- 7.14mm of the mean) and of agreement "within one scale value" ($P_{onescale}$; within +/- 21.5mm of the mean). An *a priori* minimum agreement criterion for inclusion in data analysis was set at the chance level (i.e., P_{exact} = .14; $P_{onescale}$=.39). One rater (P14) failed to meet chance probability for overall severity, roughness or strain at the P_{exact} level or for roughness at $P_{onescale}$. Based on *a priori* agreement criteria, their data were removed before further statistical analysis.

RESULTS

Marking the CAPE-V form

Participants were given no instructions other than to rate the stimuli "as they normally would," given that they were required to use a paper version of the CAPE-V form. Their markings on the CAPE-V form and scales (Figure 1), were noted and tabulated. Although most of the participants marked the VA scales with a line, two marked them with "x." In the case of "x" markings, measurement was taken from the center of the x in relation to the scale. One participant marked the line with circles, then wrote numerals between 0 and 100. One marked the line for all of the primary dimensions of voice quality and adding a numerical

rating of overall severity. Two participants both marked the line and provided a numerical rating for all of the primary dimensions of voice quality, and one provided only a numerical rating. When numerals were provided, they were accepted as the participant's judgment; however, all marks on the scales were measured and recorded. Differences between measured and numeric scores ranged between 0-8 mm for these participants and mean differences were less than 4mm (P2, *M* 3.21 [*SD* 2.42]; P15, *M* 2.88 [*SD* 1.66]; P13, *M* 2.00 [*SD* 1.54]).

Participants generally made only one mark on each scale per speaker, indicating a single judgment for each perceptual dimension based on all 5 stimuli. Only one participant (P5) marked more than one discrete location on the scale to indicate a difference in voice quality between the vowel (64mm) and sentence stimuli (74mm), notating them with "V" for vowel and "S" for sentence. This occurred for a single judgment of breathiness, which was recorded for data analysis as the average of 69mm.

Rating primary dimensions

Because participants had been asked to provide ratings as they normally would, not all chose to rate the four primary auditory-perceptual dimensions of voice quality on the CAPE-V scale. All 19 participants whose data were analyzed rated overall severity, but only 17 rated breathiness, roughness and strain. The remaining two participants reported that they do not separately rate breathiness, roughness or strain in their clinical practice, even though they may detect these separate parameters. Descriptive statistics are displayed by speaker in Table 3.

Rating pitch and loudness

Eleven participants rated pitch for all speakers; three rated pitch only if it was a concern (i.e., not 0). Five participants rated pitch inconsistently or did not indicate the nature of the abnormality (Figure 1A). In some cases, these participants indicated that pitch was not a concern by marking the line at 0; in others, they made no mark or comment regarding pitch. One participant did not rate pitch for any speaker.

Although participants were allowed to adjust the volume of presentation at the beginning of the session, two participants declined to judge the speakers' loudness because of the limitations of hearing recordings in a free field without visual or other auditory context. Ten participants rated loudness for all speakers. One rated loudness only if it was a concern, and 6 rated pitch inconsistently, with no clear pattern and not always indicating the nature of the abnormality (Figure 1B). One participant did not rate loudness at all.

Rating additional features

As shown in Figure 1C, the CAPE-V form provides two blank lines on which users can write perceptual dimensions of their choice. This option was rarely used. One participant wrote in "hypernasality" for one speaker and provided a rating. Another wrote in the following terms for six of the speakers, but did not provide ratings: "pressed, hard glottal attack, tremor, decreased prosody, unstable."

Eighteen participants either wrote comments or circled items listed on the bottom of the CAPE-V form, such as fry and asthenia (Figure 1D). One participant wrote similar comments (i.e., "fry, pressed") to the right of the scales and one participant wrote nothing on the form apart from marking a rating of overall severity.

Marking consistency

Participants' use of the markers of consistent or intermittent quality on the CAPE-V form (Figure 1E) was also varied. Seven participants circled either "C" or "I" on the form for every speaker and for every perceptual dimension. Three circled one of the letters for any perceptual dimension that was rated (i.e., not 0); one marked only "I," indicating a default value of consistency for unmarked voice quality parameters. Six participants displayed no apparent pattern to their use of these markers, and three never used them.

Descriptive statistics

Because one rater did not meet agreement criteria, statistical calculations for overall severity were based on ratings from 19 participants. Because two of the participants reported that they do not rate breathiness, roughness or strain individually in their clinical practice (and did not do so for this

TABLE 3.
Reliability and Agreement for Reduced Data Set

		Single Rater Reliability	Agreement Within 7.14mm (P_{exact})			Agreement Within 21.5mm ($P_{onescale}$)		
	N	ICC [95% CI]	M (SD)	95% CI	Range	M (SD)	95% CI	Range
Overall Severity	19	.83 [.71 - .94]	.45 (.18)	.37 - .54	.17 - .75	.90 (.12)	.84 - .95	.50 - 1.00
Breathiness	17	.76 [.61 - .91]	.45 (.14)	.37 - .53	.17 - .75	.88 (.10)	.83 - .93	.58 - 1.00
Roughness	17	.64 [.45 - .84]	.44 (.13)	.37 - .50	.25 - .75	.80 (.14)	.73 - .87	.50 - 1.00
Strain	17	.70 [.53 - .88]	.43 (.19)	.33 - .53	.17 - .75	.82 (.17)	.73 - .91	.42 - 1.00

Notes: Single rater reliability (ICC[2, 1]) calculated for absolute agreement; mean, standard deviation, confidence interval and range of P calculated for each dimension in mm, for P_{exact} and $P_{onescale}$.

study), analysis of these dimensions was based on ratings from only 17 participants.

Reliability

ICC estimates were calculated for single rater reliability based on an absolute-agreement, 2-way random-effects model. Single rater reliability was moderate-strong for overall severity and breathiness, and poor-moderate for roughness and strain, based on 95% confidence intervals, shown in Table 3.[33]

Agreement

Interrater agreement is reported as the probability that an individual listener would agree within 7.14mm (P_{exact}) and 21.5mm ($P_{onescale}$) with the group mean. Mean probability of agreement exceeded chance for all dimensions at both levels of precision. For example, Figure 2 shows the range of P_{exact} for group mean ratings of overall severity, with chance agreement indicated by the gray bar. Mean P_{exact} was poor-moderate, however, based on 95% confidence intervals, across the all dimensions of voice quality (Table 3). On the other hand, mean $P_{onescale}$ was moderate-strong, particularly for overall severity and breathiness.[33]

The relationship between years of experience evaluating voice and degree of agreement with the group mean was very weak at both levels of agreement. The linear correlation between years of experience and agreement was $r = .09$ for exact agreement and $r = .15$ for agreement within one scale value.

Range of ratings

On average, with a mean range of 48mm, ratings of overall severity varied less than breathiness (mean 49mm), roughness (mean 54mm) or strain (mean 55mm). Figure 3 displays the range of overall severity ratings for each speaker, based on data from the 19 raters.

DISCUSSION

The objective of this study was to describe how experienced voice clinicians use the CAPE-V rating scales in their daily practice and to establish the interrater reliability and agreement of their ratings of the same set of voices. Most participants used the scale in a similar way, marking the scales for at least the primary dimensions of voice quality, although they differed in the amount of detail they provided on the form. Agreement and reliability results are consistent with the literature on auditory-perceptual ratings of voice by experienced raters; that is, single rater reliability was strong in the presence of relatively weak absolute interrater agreement. The absolute range of ratings for certain speakers was

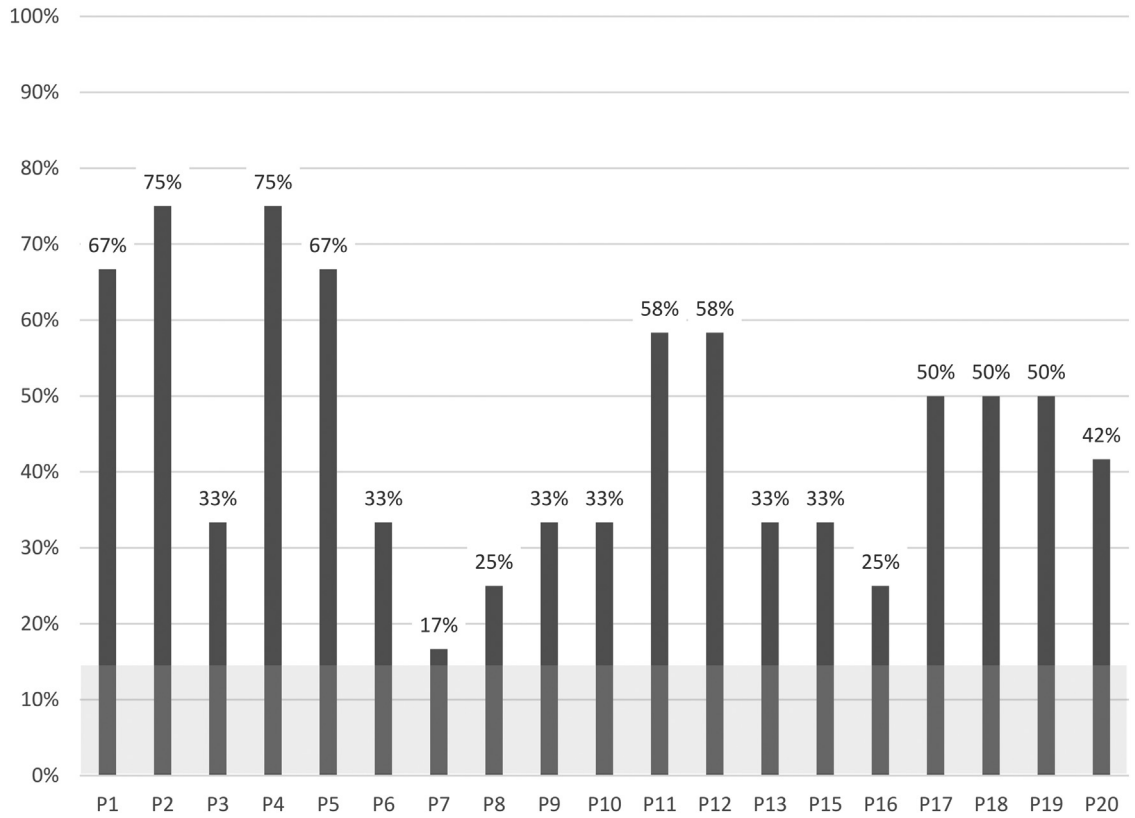

FIGURE 2. Probability of exact agreement (P_{exact}; within +/- 7.14) with group mean OS rating for 19 participants. Note: Gray area indicates probability of chance agreement (.14). (For interpretation of the referencesfig to color in this figure legend, the reader is referred to the Web version of this article.).

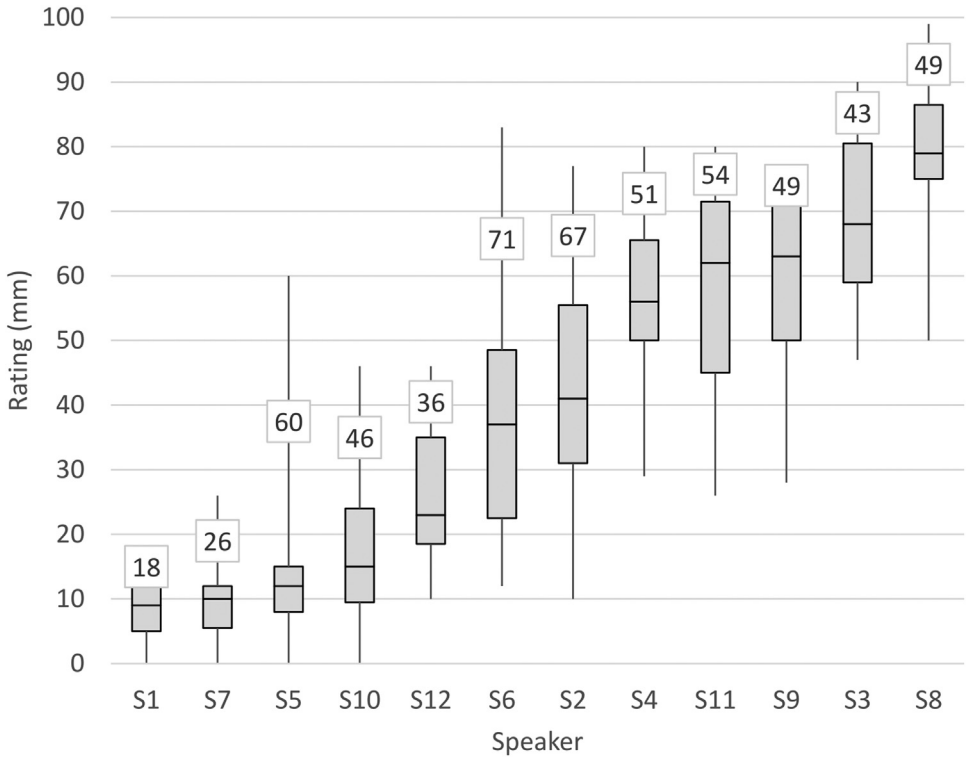

FIGURE 3. Box plot showing the labeled ranges in mm and interquartile ranges of OS ratings (n=19) for each speaker, in order of increasing median (inside line). Secondary y axis shows equally spaced severity markers (in black) as shown on the CAPE-V scales. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.).

quite high (Figure 3), in some cases spanning more than 60% of the scale, even for overall severity.

Marking the CAPE-V form

In the CAPE-V protocol, users are directed to rate overall performance (across production of sustained vowel, sentences and spontaneous speech) for each scale. If the user notes differences in quality across tasks, they are instructed to rate performance on each task separately but on the same scale line,[7] and further:

> "In the case of discrepancies across tasks, tick marks should be labeled with the task number. Tick marks reflecting vowel prolongation should be labeled #1 (see form). Tick marks reflecting running speech should be labeled #2."(p.3)[7]

In the current study, 18 of the participants marked only one location on every scale for every speaker, suggesting that in their opinions the severity of breathiness, roughness, strain and dysphonia did not vary across task for any of the speakers. One participant made a 10mm distinction in breathiness between the vowel and sentence productions in what could be referred to as the "gray area" of moderate-severe on the CAPE-V scales (i.e., 64-74mm). This participant did not otherwise distinguish performance on one speaking task from another for any other speaker (or for any other dimensions of voice quality for the speaker in question).

Although the CAPE-V does not instruct users to write in a numerical rating, those who provided both a tick mark and a number were quite accurate in their estimation of the location of the mark. The mean difference of 4mm and range of 8mm among the three participants who did both are well within the range of exact agreement set for this study. This limited evidence suggests that although these users may not have followed the CAPE-V instructions to the letter, any effect of providing a number rather than marking a point on a scale line was quite small.

The dimensions chosen for the CAPE-V scales "reflect a minimal set of clinically meaningful, perceptual voice parameters, identified by a group of expert clinicians...." (p.1).[7] Most participants in this study provided ratings for the four primary dimensions, and pitch and loudness, for all speakers; however, one participant rated only overall severity. No pattern emerged as to participants' choice to rate pitch or loudness, although two stated that they were uncomfortable judging loudness based on hearing recordings presented in a free field. It may be that some participants resist rating pitch and loudness in favor of relying on measures of their acoustic correlates (i.e., speaking fundamental frequency; dB SPL), but it seems more likely that they choose to mark the form only if they recognize some abnormality in these dimensions. However, not every participant who rated pitch or loudness described the nature of

the deviance, suggesting that they may have had an internalized understanding of the placement of tick marks such that they would expect to remember whether pitch or loudness were increased or decreased.

Most participants used the C/I markers of consistent or intermittent in a methodical way, either for all speakers, or for only those with intermittent voice quality issues. The six participants who showed no clear pattern of use of these markers may have had an internalized method that was not identifiable to the examiner. For the participants who never used the markers, it is impossible to know whether they detected no inconsistencies in quality, pitch or loudness, or whether it is just not their practice to mark them on the CAPE-V form.

Most of the participants either circled or wrote in some other aspect of voice quality already listed on the form, such as "vocal fry, asthenia, tremor." Only two of the participants used the extra blank scales, however, and only one provided a rating on that scale (of hypernasality). Given that the other participant using the blank scales wrote in a descriptor without providing a rating, it seems that the blank scale may not serve much of a purpose to experienced voice clinicians. Qualitative analysis of interview data from these participants is ongoing and should provide greater context for these observations.

Single rater reliability

Single rater reliability for this study was comparable to similar research using the CAPE-V protocol and reporting single-measures ICCs. In the validation study of the CAPE-V, ICC (2,1) coefficients for 21 expert raters were moderate (i.e., overall severity r = .76; breathiness r = .60; roughness r = .62; strain r = .56); confidence intervals were not reported.[3] Ratings were based on conversational speech samples from 59 speakers (37 with dysphonia), not CAPE-V sentences. Single rater reliability for three expert raters judging only overall severity was also moderate (ICC [2,1] = .75, 95% CI [.64-.83]) based on six CAPE-V sentences and two vowels produced by 56 speakers.[38] In that study, male and female voices were evaluated separately and with a "moderately severe" anchor sample of the same sex. Given that participants in the current study heard a mix of male and female speakers with no such anchor sample, the moderate level of single rater reliability found for overall severity (ICC[2,1] 83%) might be something of an overperformance. It may represent the best interrater reliability to be expected by experienced clinicians using a VA scale with text markers to rate a range of dysphonic voices.

Interrater agreement

Two levels of interrater agreement were set for this study. Although assigning a narrow range of absolute ratings to an anchor sample is ideal for training purposes, previous evidence suggested that examining a wider range of agreement, within one scale value of a group mean, would be necessary to capture the variability among expert raters. Auditory anchors might be accompanied by a wide range of expert ratings that agreed within +/- 21.5mm and specifying a subset of those ratings within +/- 7.14mm to get a sense of confidence about the mean/median rating. At the higher level of precision, mean P_{exact} was slightly higher for all primary dimensions of voice quality in the current study than for results reported by Eadie and Kapsner-Smith (2011). Mean P_{exact} among all included participants in the current study exceeded chance probability of 14%; however, it was still relatively weak across dimensions, with none achieving 50% probability of agreement with the group mean (Table 3).

Mean probability of agreement "within one scale value," or +/- 21.5mm, was moderate to strong, based on 95% confidence intervals for all primary dimensions of voice quality, exceeding reports of $P_{onescale}$ by Eadie and Kapsner-Smith for overall severity and strain.[4] Individual levels of agreement varied considerably, but at least one participant provided ratings within one scale value of the group mean for all 12 speakers (i.e., 100% agreement) for each of the primary voice quality dimensions. However, setting agreement at +/- 21.5mm meant that ratings as far apart as 43mm (i.e., 43% of the scale length) might agree. A difference of this size could span two or even all three labeled categories of severity, depending on the user's interpretation of the text markers shown in Figure 3. For this reason, the interaction of acoustic metrics of voice samples on auditory-perceptual ratings should be investigated, specifically for the CAPE-V.

The criterion of chance agreement at both levels for inclusion of participant data in statistical analysis was not intended to present a challenge to experienced voice clinicians; however, it did require removal of one participant's data from analysis of range, agreement and reliability. The participant whose data did not meet agreement criteria for statistical analysis, acknowledged that although they had more than five years' experience evaluating voice, that experience was limited to about one assessment every two weeks. They also described their voice practice as mostly composed of singers. A clinician whose internal standard for dysphonic voice is based on singers' voices may perceive voices differently from one whose referent is the speaking voice.[39]

The weak relationship between years of experience and agreement with the group mean is somewhat surprising. Evidence of an effect of experience specifically on interrater agreement is limited, but Eadie and Kapsner-Smith reported similar likelihood of agreement at the $P_{onescale}$ level for inexperienced (72%) vs. experienced (75%) raters of overall severity (although they differed at the P_{exact} level by 14%).[4] It may be that years of experience is a poor proxy for the amount of time actually spent evaluating voices. Future research should examine the association between relative experience and interrater agreement and reliability for judgments of dysphonia.

Ranges of ratings

Despite removing data from one outlier participant in the current study, and despite the effect of that removal on

interrater agreement, ranges of overall severity, breathiness, roughness and strain ratings were sometimes extremely large. For example, mean ranges of roughness and strain were 54mm and 55mm, respectively, representing 54% and 55% of the length of these scales, even larger than the range represented by agreement within one scale value. Even without the complication of the gradations of severity labeled below the scales, these ranges represent problematic levels of disagreement among experienced clinicians, although they are similar to the findings reported by Walden.[11]

When reviewing ratings obtained from a hybrid VA scale such as the CAPE-V, it is critical to consider the degree to which a given range of ratings may encompass more than one category. Given that the presence and location of subscale text markers may affect ratings of voice quality, it is important that the same scale be used when comparing intra- and interrater judgments.[13]

Irrespective of their actual placement, textual markers serve to partition VA scales into categories, and clinicians routinely use descriptors of "mild, moderate, severe" to describe voice quality, as the corpus of stimuli used for this study showed. Depending on the placement of the text markers and the clinician describing the voice, a voice rated as low as 20mm or as high as 70mm might be described as "moderately" breathy (Figure 3). Ratings of overall severity for Speaker 6 from the current study show an even greater range than this, with a low rating of 12, which may be considered normal (or arguably mild), and a high rating of 82, which is clearly severe. Voice samples from such a speaker are unlikely to provide good prototypes of any one dimension of voice quality, but their existence serves to highlight the wide variability among voice clinicians. If experienced voice clinicians are this divergent in their criteria for rating overall severity, which generally exhibits the highest interrater reliability of the dimensions rated on the CAPE-V, new clinicians and clinicians new to rating voice quality should be made aware of this variability.

The instructions for the CAPE-V indicate that the labeled general regions below each scale "represent gradations in severity, rather than discrete points."[7] Still, at some point, there may be a JND at which point "mild" becomes "moderate" and "moderate" becomes "severe." There may also be unlabeled areas beneath each scale that signify a "mild-moderate" or "moderate-severe" degree of deviance to some or all users of the CAPE-V. Examining where experienced clinicians apply these boundaries would provide guidance to learners and infrequent raters of auditory-perceptual qualities of voice. Future research should elicit narrative descriptions of voice quality and degree of severity that address the location on the scale at which they mark these transitions.

Finally, a quick view of Figure 3 shows relatively narrow interquartile (IQ; middle 50% of ratings) ranges for speakers in the "mild" (i.e., S1, S5, S7, S10) and "severe" ranges (i.e., S8), based on median ratings of overall severity. The wider IQ ranges for speakers with "moderate" overall severity are consistent with the reduced agreement among voice quality ratings in the "mid-range" frequently reported in the literature.[28] The wider IQ ranges in the middle of the scales suggest that these experienced clinicians had as much difficulty evaluating moderately dysphonic speakers as previous research has indicated.[4,37,40]

External standards

For voice clinicians, access to some external standard of auditory-perceptual characteristics of voice quality is an essential tool of evaluation, but there is currently no single standard set of recordings to serve even as a training tool.[41] One of the long-term goals of this research is to improve the psychometric and ecological validity of ratings that accompany freely available voice recordings such as the UWVDS and PVQD.[14] The results of this study indicate that a single mean or median rating of any dimension of voice quality fails to capture the variability among experienced clinicians who listen to dysphonic voices on a daily basis. Providing a median and IQ range of ratings from a group of experienced clinicians, rather than a single rating from one or a few, would allow users of a potential external standard or auditory anchor to recognize the variability inherent in evaluating some voice quality dimensions for some speakers. Relatedly, although individual voice quality dimensions measured on the CAPE-V protocol are multidimensional, absolute inter-rater agreement is greater for overall severity and breathiness than it is for roughness and strain. The IQ ranges of overall severity and breathiness ratings were relatively small in the current study (i.e., 30mm or less), whereas some of the IQ ranges of roughness and strain ratings were more than 35 mm. Of course, raters may agree substantially on one dimension of voice quality, such as breathiness, while varying considerably on another.

Training cannot always address composite voice abnormalities where features are variable. For this reason, it might be worthwhile to identify voices with narrow IQ ranges of ratings in all dimensions for a first phase in training auditory-perceptual evaluation of voice. Once raters could achieve a level of agreement within the middle 50% of "expert" ratings for each dimension of voice quality, a second phase of training could be introduced. In the second phase, voices might be chosen to represent gradations of severity based on a median expert rating of a single dimension, regardless of variability around the median for the other voice quality dimensions. A third phase might include complex voices for which the IQ ranges might be relatively wide, but that represented clinically relevant external standards.

Training to this type of standard might force the rater to find the "best fit" rather than providing a valid index of voice quality, at least for the duration of training. However, standard auditory-perceptual referents of voice quality, along with rating information such as is reported here, are essential tools for exposing new clinicians to an array of dysphonic voices and for recalibrating experienced clinicians who may not encounter them every day. When descriptive details beyond the primary dimensions of voice quality are necessary to capture essential characteristics of a voice,

clinicians will remain free to provide them on the CAPE-V form or elsewhere.

Limitations

This study is part of a project investigating the use of the CAPE-V among experienced voice clinicians. Several aspects of the study limit its generalizability beyond this population. First, participants heard recordings from only 12 speakers. This was largely in the interest of the participants' time. The bulk of the interview sessions was dedicated to collecting narrative descriptions of how the participants used the CAPE-V protocol rather than simply obtaining quantitative data. Sessions ran for up to 2 hours, even with this small number of recordings.

Second, only a portion of the standard CAPE-V recording protocol for each speaker was presented; two of the CAPE-V sentences (one including all English vowels and one specifically sampling nasals) and the spontaneous speech sample were not available. Although it is possible that the missing sentences would not have had any effect on ratings of voice quality for the speakers in this study, it is likely that the addition of spontaneous speech samples would have. (One participant mentioned this several times.)

Third, although the published instructions for use of the CAPE-V[1] were followed as closely as possible, participants were directed specifically to use the CAPE-V the way they normally would. This meant, for example, that some participants requested multiple repetitions of recordings, while others requested none. Some participants did not rate the dimensions of breathiness, roughness or strain at all. To the degree possible, the context for obtaining data in this study was controlled, but given the goal of examining "real-world" use of the CAPE-V instrument, these limitations must be taken into account.

Participants in this study were recruited from four geographical areas in the United States but may not have represented even typical voice clinicians. Some of them were employed at elite voice clinics, and some were current or future academics. Finally, the degree to which a clinician is "expert" is difficult to quantify, apart from attempting to prescribe a minimum of current and past exposure to voice disorders. Several years of experience doing things very differently from one's colleagues would still count as "experience" based on this criterion. The relative contribution of individual experience to interrater reliability and agreement could instead be indexed in multiple ways (e.g., years of experience; typical number of clinical hours spent evaluating voice; current proportion of evaluations on caseload). Despite these limitations, the results of this study reflect current practice among a group of clinicians who think about voice every day.

CONCLUSION

This study supports the idea that establishing external standard auditory anchors for clinical use may require the use of a range of ratings, rather than a single numerical value. Results indicated differential use of the components of the CAPE-V form and scales among experienced voice clinicians evaluating the presence and degree of severity of dysphonia. Data also revealed frequently categorical variability among ratings of all of the primary CAPE-V dimensions of voice quality that may complicate the clinical description of a voice as mildly, moderately or severely dysphonic. These results are concerning in the context of attempting to train clinicians to a standard, and suggest further examination of: 1) how experienced voice clinicians interpret the instructions on the CAPE-V; 2) how they would describe dysphonic stimuli in words; and 3) how we might identify or create auditory anchor samples on which greater consensus could be found.

Investigation of the relationship between quantitative ratings and qualitative descriptions of the stimuli used in this study is ongoing. Detailed descriptions of clinicians' thought processes as they evaluated specific stimuli presented in this study could shed light on some of the differences reported here. For example, some clinicians might report that they rate their overall impression of a voice first, then listen specifically for roughness, breathiness or strain. Others might rate voice in person during a session, listening only once and characterizing the voice only by its most salient or severe dimension. Knowing how experienced clinicians use the CAPE-V protocol and scales will help to clarify whether a revision of the instrument itself is in order or a re-introduction of the current protocol (with the noted corrections) will suffice. Relatedly, comparing verbal descriptions of stimuli will establish usage patterns that may exist among this set of experienced clinicians and provide some insight into their auditory-perceptual ratings of voice more generally.

Whether we identify or create auditory anchors, achieving better consensus on the presence and severity of roughness, breathiness, strain and overall deviance from typical voice for a set of recordings is important. A bank of recordings, rated and described by experienced voice clinicians under optimal listening conditions might be a way to start. A larger set of "expert" CAPE-V ratings, showing the range for each UWVDS or PVQD recording, could also be built to supplement these existing online resources so that any clinician could benefit from this program of research.

Acknowledgments

This work was funded by a 2018 American-Speech-Language-Hearing Foundation New Investigators Research Grant. The author gratefully acknowledges the Center for Laryngeal Surgery and Voice Rehabilitation at the Massachusetts General Hospital for sharing the de-identified data that was used to create the stimuli for this study. The author is also grateful to Dr. Gabriel Cler (University of Washington) and Dr. Philip Doyle (Stanford University) for their insightful comments on the

manuscript, and Alissa Loffreno for her tireless work on this project. The author thanks the participants in this study for sharing their time, energy, and expertise. Portions of the data reported here were presented at the Fall Voice conference in Dallas, TX (2019) and virtual (2020); and at the virtual 50th Annual Voice Foundation Symposium (2021).

REFERENCES

1. Kempster GB, Gerratt BR, Verdolini Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
2. Karnell M, Melton S, Childes J, et al. 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.
3. 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.
4. 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.
5. Helou L, Solomon N, Henry L, et al. The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. *Am J Speech Lang Pathol*. 2010;19:248–258.
6. Nagle KF. Challenges to CAPE-V as a standard. *Perspect ASHA Spec Interest Groups SIG*. 2016;1:47–53. https://doi.org/10.1044/persp1.SIG3.47. *3*.
7. ASHA. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Division 3, Voice and Voice Disorders. Published online 2009. Accessed October 13, 2019. https://www.asha.org/Form/CAPE-V-Success/.
8. Lodhavia A, Kempster GB. *Fidelity to the CAPE-V*. Denver, CO: ASHA; 2015. Poster presented at:November 13.
9. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103–115.
10. Kreiman J, Gerratt BR. Comparing two methods for reducing variability in voice quality measurements. *J Speech Lang Hear Res*. 2011;54:803–812. https://doi.org/10.1044/1092-4388(2010/10-0083.
11. Walden PR. Perceptual voice qualities database (PVQD): database characteristics. *J Voice*. 2020;0. https://doi.org/10.1016/j.jvoice.2020.10.001.
12. Hirano M. GRBAS" scale for evaluating the hoarse voice & frequency range of phonation. In: Hirano, ed. *Clinical Examination of Voice. Vol 5. Disorders of Human Communication*. Springer-Verlag/Wien; 1981:83–84. 88-89.
13. Nagle KF, Helou LB, Solomon NP, et al. Does the presence or location of graphic markers affect untrained listeners' ratings of severity of dysphonia? *J Voice*. 2014;28:469–475. https://doi.org/10.1016/j.jvoice.2013.12.011.
14. Walden P. Perceptual Voice Qualities Database (PVQD), Mendeley. Published online October 9, 2020. Accessed October 20, 2020. https://data.mendeley.com/datasets/9dz247gnyb/3.
15. 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.
16. Eddins DA, Anand S, Camacho A, et al. Modeling of breathy voice quality using pitch-strength estimates. *J Voice*. 2016;30:774.e1–774.e7. https://doi.org/10.1016/j.jvoice.2015.11.016.
17. Eddins DA, Shrivastav R. Psychometric properties associated with perceived vocal roughness using a matching task. *J Acoust Soc Am*. 2013;134:EL294–EL300. https://doi.org/10.1121/1.4819183.
18. Anand S, Kopf LM, Shrivastav R, et al. Objective indices of perceived vocal strain. *J Voice*. 2019;33:838–845. https://doi.org/10.1016/j.jvoice.2018.06.005.
19. Eddins DA, Anand S, Lang A, et al. Developing clinically relevant scales of breathy and rough voice quality. *J Voice*. 2021;35:663.e9–663.e16. https://doi.org/10.1016/j.jvoice.2019.12.021.
20. Stevens S. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. Wiley; 1975.
21. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *J Speech Lang Hear Res*. 2002;45:1088–1096. https://doi.org/10.1044/1092-4388(2002/087.
22. Schiavetti N, Sacco PR, Metz DE, et al. Direct magnitude estimation and interval scaling of stuttering severity. *J Speech Hear Res*. 1983;26:568–573.
23. Toner MA, Emanuel FW. Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *J Speech Hear Res*. 1989;32:78–82.
24. Stevens SS, Galanter EH. Ratio scales and category scales for a dozen perceptual continua. *J Exp Psychol*. 1957;54:377–411. https://doi.org/10.1037/h0043680.
25. Bunton K, Kent RD, Duffy JR, et al. Listener agreement for auditory-perceptual ratings of dysarthria. *J Speech Lang Hear Res*. 2007;50:1481–1495.
26. Yiu E, Ng C. Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clin Linguist Phon*. 2004;18:211–229.
27. Wilson DK. *Voice Problems of Children*. 3rd ed. Williams & Wilkins; 1987.
28. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.
29. Kreiman J, Gerratt BR, Precoda K, et al. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35:512–520.
30. Eadie T, Nicolici C, Baylor C, et al. Effect of experience on judgments of adductor spasmodic dysphonia. *Ann Otol Rhinol Laryngol*. 2007;116:695–701.
31. Eadie TL, Kapsner M, Rosenzweig J, et al. The role of experience on judgments of dysphonia. *J Voice*. 2010;24:564–573. https://doi.org/10.1016/j.jvoice.2008.12.005. S0892-1997(08)00209-9 [pii][doi].
32. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
33. Portney L, Watkins M. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Prentice Hall Health; 2000.
34. Wuyts F, De Bodt M, Van de Heyning P. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508–517.
35. Ghio A, Dufour S, Wengler A, et al. Perceptual evaluation of dysphonic voices: can a training protocol lead to the development of perceptual categories? *J Voice*. 2015;29:304–311. https://doi.org/10.1016/j.jvoice.2014.07.006.
36. Awan SN, Roy N, Jetté ME, et al. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the CAPE-V. *Clin Linguist Phon*. 2010;24:742–758. https://doi.org/10.3109/02699206.2010.492446.
37. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am*. 1998;104:1598–1608.
38. Awan SN, Solomon NP, Helou LB, et al. Spectral-cepstral estimation of dysphonia severity: external validation. *Ann Otol Rhinol Laryngol*. 2013;122:40–48.
39. Oates JM, Bain B, Davis P, et al. Development of an auditory-perceptual rating instrument for the operatic singing voice. *J Voice*. 2006;20:71–81. https://doi.org/10.1016/j.jvoice.2005.01.006.
40. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45:111–126. https://doi.org/10.1044/1092-4388(2002/009.
41. Iwarsson J, Reinholt Petersen N. Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *J Voice*. 2012;26:304–312. https://doi.org/10.1016/j.jvoice.2011.06.003.