Establishing Validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Richard I. Zraick
University of Arkansas for Medical Sciences, Little Rock

Gail B. Kempster
Rush University, Chicago, IL

Nadine P. Connor
Susan Thibeault
University of Wisconsin—Madison

Bernice K. Klaben
University of Cincinnati Physicians, West Chester, OH

Zoran Bursac
Carol R. Thrush
University of Arkansas for Medical Sciences

Leslie E. Glaze
University of Minnesota, Minneapolis

Purpose: The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed to provide a protocol and form for clinicians to use when assessing the voice quality of adults with voice disorders (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kramer, & Hillman, 2009). This study examined the reliability and the empirical validity of the CAPE-V when used by experienced voice clinicians judging normal and disordered voices.
Method: The validity of the CAPE-V was examined in 2 ways. First, we compared judgments made by 21 raters of 22 normal and 37 disordered voices using the CAPE-V and the GRBAS (grade, roughness, breathiness, asthenia, strain; see Hirano, 1981) scales. Second, we compared our raters' judgments of overall severity to a priori consensus judgments of severity for the 59 voices.
Results: Intrarater reliability coefficients for the CAPE-V ranged from .82 for breathiness to .35 for strain; interrater reliability ranged from .76 for overall severity to .28 for pitch.
Conclusions: Although both CAPE-V and GRBAS reliability coefficients varied across raters and parameters, this study reports slightly improved rater reliability using the CAPE-V to make perceptual judgments of voice quality in comparison to the GRBAS scale. The results provide evidence for the empirical (concurrent) validity of the CAPE-V.

Key Words: Consensus Auditory-Perceptual Evaluation of Voice, CAPE-V, voice, voice assessment

Disorders of voice present as deviations in voice quality and are thus by nature auditory-perceptual phenomena. That is, a listener recognizes that a particular voice sounds unpleasant or seems inadequate relative to what is perceived to be "normal" by that listener. As noted by Carding, Wilson, MacKenzie, and Deary (2009) in a recent review of voice outcomes, it is a perceived disruption in voice quality that leads individuals to seek treatment. Furthermore, voice quality assessments have been shown to influence the direction and course of voice therapy (Behrman,

2005) and to determine whether patients have improved with treatment (Carding et al., 2009). In other words, how the voice *sounds* does matter. Accordingly, it is important to accurately and consistently describe and quantify the *quality* of a person's voice.

To be a useful voice outcome measure for clinical and research purposes, any agreed-upon assessment method should be reliable, valid, and responsive to change (Carding et al., 2009). Speech scientists have applied advanced acoustic synthesis and filtering techniques in attempts to develop a methodology for reliable measurement of vocal quality (Gerratt & Kreiman, 2001; Shrivastav, Sapienza, & Nandur, 2005). However, this work is not currently applicable to clinical settings. To date, no single method of auditory-perceptual voice analysis has achieved these criteria, nor has any instrument been used consistently by the voice community. Indeed, this dilemma has confounded communication of clinical and research findings.

Because auditory-perceptual measures of voice would serve as reasonable, intuitive, and tangible voice outcomes, it is unfortunate that a standardized method for implementing voice quality judgments has been slow to develop. This endeavor is difficult because of the many psychophysical scaling issues that influence the task, as long considered in the psychology literature (Marks & Algom, 1998; Stevens, 1975; see Kent, 1996, for a detailed review). In their well-known tutorial, Kreiman, Gerratt, Kempster, Erman, and Berke (1993) reviewed 57 different articles selected from the literature that used various approaches to auditory-perceptual analysis of voice. Among these approaches, the GRBAS scale, introduced officially in the English-speaking world by Hirano (1981), has been widely used for judging disordered voice quality (Carding et al., 2009). Each parameter on the GRBAS scale represents a dimension of phonation: G (grade) represents the degree of overall voice abnormality, R represents roughness, B represents breathiness, A represents asthenia (weakness), and S represents strain. The GRBAS uses a 4-point Likert scale of 0 (*normal*) to 3 (*extreme*) for all five parameters. However, the GRBAS scale does not offer a specific protocol for administration and does not provide guidelines for analysis. Also, an ordinal scale, like that used in the GRBAS, does not allow parametric statistical analysis. These and other issues noted by Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kramer, and Hillman (2009) as limitations of the GRBAS were found to influence the reliability of voice quality assessments in studies of auditory-perceptual analysis (Gorodetsky, Amir, & Yarom, 1992). Such issues include lack of clarity regarding the amount and type of training, the possible influence of task order effects, and variability of listening samples (Kreiman, Gerratt, & Ito, 2007; Kreiman et al., 1993). Thus, the GRBAS, despite its wide use, may not result in reliable or valid voice quality judgments and thus may not provide optimal voice outcome measures for clinical or research purposes.

To address these concerns, a new tool for auditory-perceptual voice measurement was developed that uses continuous scaling, involves a variety of speaking tasks and voice contexts, and provides a detailed protocol for voice sample recording and data analysis. The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed under the auspices of Special Interest Division 3 (Voice and Voice Disorders) of the American Speech-Language-Hearing Association (ASHA). A description of this consensus effort, the rationale that underlies the test items, the recording and analysis procedures, and a reproduction of the form can be found in Kempster et al. (2009). The CAPE-V uses continuous visual analog scales for judgments of six parameters of voice: overall severity, roughness, breathiness, strain, pitch, and loudness. When using the CAPE-V, the clinician places a vertical tick mark on a 100-mm horizontal line to denote the severity of the disorder, with a higher value indicating greater severity. Thus, continuous interval data between 0 and 100 can be derived for each aspect of voice quality and applied to statistical analysis where appropriate. The CAPE-V also allows the clinician to note other voice features for a particular patient, as needed.

A few studies have examined the reliability and validity of the CAPE-V for measuring constructs of voice quality disorders. In the context of a larger study, Karnell et al. (2007) reported results from four voice clinicians from the same institution who judged voice samples using the CAPE-V and the GRBAS. They achieved high correspondence between these rating methods and good intra- and interrater reliability on the CAPE-V (Karnell et al., 2007). Similarly, a recent study examining CAPE-V reliability for disordered pediatric voices found excellent agreement within and across three raters from the same setting (Kelchner et al., 2010). However, other data are scarce. Thus, a larger study of CAPE-V reliability and validity that includes experienced listeners drawn from voice centers across the United States appears warranted. This will help determine the reliability of the CAPE-V made by raters representing diverse training backgrounds, geographical regions, and clinical settings.

The reliability of a rating scale is the degree to which judgments derived from that scale are dependable or consistent within a rater or across raters on repeated administrations. Validity is concerned with the extent to which a scale's scores can be interpreted as representative of a particular underlying construct (Carding et al., 2009; Carmines & Zeller, 1979; Cook & Beckman, 2006; Cronbach & Meehl, 1955; DeVon et al., 2007; Kelly, O'Malley, Kallen, & Ford, 2005; Sechrest, 2005; Shadish, Cook, & Campbell, 2002). Types of validity, such as content, face, construct, criterion, empirical, convergent, and predictive, can be defined and assessed when new instruments or scales are developed (DeVon et al., 2007).

In their seminal work, Cronbach and Meehl (1955) described four forms of validity: predictive, concurrent, content, and construct. The first two are examples of criterion-based validity, which compare test results to an established standard. Predictive validity is the degree to which a new probe predicts performance on a different test, such as using CAPE-V results to predict a patient's anticipated quality of life (Karnell et al., 2007). In contrast, concurrent validity reflects the ability of new test items to replace a comparable measure. For example, to what extent might judgments using the CAPE-V correspond to those made using the GRBAS scale? The third type, content validity, reflects the adequacy of the test items in representing the *universal* underlying measurement sample. For example, do the parameters assessed in the CAPE-V sufficiently capture the multidimensional attributes of voice quality present in connected speech? Finally, construct validity is inferred indirectly, using a variety of test items to define a measurement sample when there is no established criterion available. Construct validity continues to challenge voice scientists and clinicians, who have struggled to identify parameters that represent the auditory-perceptual features of voice quality (Kempster et al., 2009; Kreiman & Gerratt, 1998).

In a recent review, Sechrest (2005) noted that establishing validity of a scale may depend on the manner in which an instrument is used. Whereas establishing reliability of measurement can be a straightforward process, establishing the validity of a test is more complex. As Sechrest clarifies, "The crux of the matter lies in Messick's assertion that 'Validity is not a property of the test or assessment as such, but rather of the *meaning* [italics added] of the test scores.' It is not measures that are valid, but the scores that they yield and the interpretation we make of them" (Messick, 1995, p. 741, as cited by Sechrest, 2005).

Sechrest (2005) described construct validity similarly to Cronbach and Meehl (1955) as the "extent to which we can legitimately claim that a measure reflects variability in the construct it purports to measure" (p. 1586). This process requires more evidence than establishing empirical, face, content, or criterion validity. Due to its complexity, construct validity may only emerge over time as an instrument is used and studied from multiple perspectives (Sechrest, 2005). Like concurrent validity (Cronbach & Meehl, 1955), empirical validity also compares a new instrument with another instrument that in theory measures the same construct. This determination can be one step in the process of estimating the construct validity of a new scale (Kelly et al., 2005; Sechrest, 2005).

Accordingly, in the current study, empirical validity is defined as the correspondence of CAPE-V and GRBAS judgments, where possible. The rationale for this comparison is that the GRBAS is a well-used and well-studied instrument for judging voice quality that is based on the same underlying constructs as the CAPE-V. Furthermore, the GRBAS has been shown previously to be a reliable instrument for the assessment of voice quality disturbances (Dejonckere, Obbens, de Moor, & Wieneke, 1993; Karnell et al., 2007; Webb et al., 2004). In addition, when both instruments are used to rate the same voice samples, GRBAS scores have been shown to correlate strongly with those of the CAPE-V (Karnell et al., 2007).

The purpose of the current study was twofold: first, to examine intrarater and interrater reliability of experienced voice clinicians' judgments of voice quality using the CAPE-V and GRBAS, and second, to establish the empirical validity of the CAPE-V by assessing relationships between the two scales. In contrast to most prior studies that examined either the CAPE-V or the GRBAS scale (De Bodt, Wuyts, Van de Heyning, & Croux, 1997; Kelchner et al., 2010), we elicited scores for both instruments using the same voice samples measured in two separate sessions, one each for the CAPE-V and GRBAS. Also, because prior research has shown that experienced voice clinicians may have highly flexible perceptual strategies (Gorodetsky et al., 1992; Lazarus, 2009), we recruited experienced clinicians from multiple voice centers across the United States, to avoid potential institutional bias that could arise from using listeners from the same institution. Our three research questions were:

1. What are the intrarater and interrater reliability judgments of the GRBAS and CAPE-V made by experienced raters?
2. How do the CAPE-V measures compare to judgments obtained on the same stimuli using the GRBAS?
3. How do experienced raters' GRBAS grade judgments compare to unanimous a priori consensus severity judgments?

Method

Human Subjects Protection

This investigation was reviewed and approved by the institutional review boards of the University of Arkansas for Medical Sciences (UAMS) and the University of Wisconsin—Madison, as well as the human subjects approval process at the Blaine Block Institute for Voice Analysis and Rehabilitation (Dayton, OH).

Voice Stimuli

Dysphonic voice samples. Two hundred dysphonic voice samples were identified from an existing clinical voice database maintained at the Blaine Block Institute for Voice Analysis and Rehabilitation. All recordings included speech productions required by the published CAPE-V protocol (Kempster et al., 2009), including sustained /a/ and /i/ vowels, six sentence repetitions, and a brief sample of conversational speech in response to a consistent question prompt. The same recording procedure was used to obtain all samples, using the KayPentax Computerized Speech Lab Model 4500 with a sampling rate of 22 kHz (Delyiski, Shaw, & Evans, 2005). A headset microphone (AKG Model C420) was used, maintained at a distance of 5 cm from the speaker's mouth. Ambient room noise was minimal.

The dysphonic voice samples consisted of 62 male voices and 138 female voices representing a range of disorders, ages, and severity levels. Two listeners screened all 200 dysphonic voice samples; 14 voice samples were rejected due to technical problems, leaving 186 stimuli. These 186 voices were judged independently for severity by five a priori experienced raters using a 4-point Likert scale (1 = *normal,* 2 = mild *dysphonia,* 3 = *moderate dysphonia,* and 4 = *severe dysphonia*). The experienced raters who made these judgments met the same inclusion criteria as all raters used in the study (see description below). Thirty-seven dysphonic samples achieved unanimous agreement across a priori raters (seven male voices and 30 female voices; male mean age = 53.5 years, range = 22–54 years; female mean age = 52.5 years, range = 18–86 years). Of the 37, all five raters judged 13 to be mildly impaired, 11 moderately impaired, and 13 severely impaired. None were judged as normal. The dysphonic voices represented a range of pathologies, using the scheme described in the Classification Manual for Voice Disorders—I

(Verdolini, Rosen, & Branski, 2006) and included structural, inflammatory, neurological, and other pathologies.

Normal voice samples. Twenty-two normal voice samples (six male, 16 female; male mean age = 40.5 years, range = 32–60 years; female mean age = 25.5 years; range = 18–38 years) were obtained from healthy volunteers with no history of vocal pathology from the University of Arkansas at Little Rock. The voices were confirmed perceptually by the first author as having normal quality. The recordings followed the CAPE-V protocol (Kempster et al., 2009). All recordings were made as described above except a handheld microphone (Shure Model SM48) was used, with a consistent 5-cm mouth-to-microphone distance.

Listening Disc

Seventy-four voice samples, each separated by 3 s of silence, were mastered onto a CD. The recordings were not normalized for intensity, and noise reduction was not applied. The 74 samples included (a) the 59 voices (37 dysphonic and 22 normal), (b) 11 repeated voices (seven dysphonic and four normal) chosen randomly to assess intrarater reliability, and (c) four voices (two dysphonic and two normal) used for task familiarization. The four familiarization voices were mastered as Tracks 1–4, with a male and female normal speaker presented first, followed by the two dysphonic voices. Using a 4-point Likert scale (1 = *normal,* 2 = *mild dysphonia,* 3 = *moderate dysphonia,* and 4 = *severe dysphonia*), the first author judged one of the dysphonic voices as mildly dysphonic and the second voice as severely dysphonic. Listeners were instructed to "Listen first to these four tracks to become accustomed to the kind of voice samples on the remaining tracks." The remaining 70 test voices were mastered as Tracks 5–74. No speaker identification information was provided for any track.

Raters

Twenty-one ASHA-certified speech-language pathologists with expertise in the assessment and treatment of persons with voice disorders were recruited as raters from a posting on ASHA's Special Interest Division 3 (Voice and Voice Disorders) e-mail list (sid3voice@list.healthcare.uiowa.edu). Selection criteria were as follows: (a) more than 5 years of experience working with voice disordered patients; (b) a caseload of voice clients seen weekly; (c) native speakers of English; (d) no history of impairments in cognition, speech, voice, language, hearing, or vision; (e) familiarity with both the CAPE-V and GRBAS scales; and (f) willingness to complete the judging tasks within a 72-hr time frame. Sixteen women and five men participated, representing 17 different facilities. Participants reported an average of 13 years of experience. Participants received a complete study packet that included the CD and a rater information form with queries about credentials, institutional affiliation, years of clinical experience, dates and duration of listening sessions, and listening equipment used. Written instructions were provided for listening and judging, and a sufficient number of CAPE-V and GRBAS forms were included.

Judging Procedure

In preparation for the judging task, all raters listened to Tracks 1–4 to become familiar with the recordings. The raters had been randomly assigned to one of two counterbalanced listening conditions: Group A clinicians (*n* =11) judged the voices using the GRBAS first, followed by the CAPE-V in the second session; Group B clinicians (*n* = 10) judged voices using the CAPE-V in the first session, followed by the GRBAS in the second session. All 21 raters listened to each of the voice samples in two sessions separated by 48–72 hr.

Raters were asked to make judgments based on the conversational speech sample. They listened to the voice samples in a free-field environment they judged to be free of potential distraction and excessive ambient noise. They were allowed to set a playback volume that was personally comfortable and were instructed to take a short (5–10-min) break after Track 36 and then to resume the session. Raters were allowed to listen to a voice sample more than once in order to make a judgment. The mean time to complete the CAPE-V session was 1.75 hr (range = 1.25–2.5 hr). The mean time to complete the GRBAS session was 1.25 hr (range = 0.5–2.0 hr). Raters returned the data forms and the CD to the first author. Data were entered into a spreadsheet for statistical analysis using SAS Version 9.

Results

Intrarater Reliability

Intrarater reliability analyses were based on repeated values for 11 stimuli. Table 1 shows the average intrarater reliability coefficients (Pearson's *r*) for each of the six CAPE-V scales, as well as the highest and lowest individual intrarater reliability coefficients. Intrarater reliability was highest for breathiness (r = .82) and lowest for strain (r = .35). Intrarater reliability can also be evaluated by assessing the number of raters whose intrarater reliability was considered good at $r > .70$, as shown in Table 1. At least 14 out of 21 raters achieved a reliability value of greater than .70 on three of the six CAPE-V scales: breathiness, roughness, and pitch. Three other CAPE-V scales—strain, overall severity, and loudness—proved more difficult for raters to use reliably in judging repeated stimuli. No rater demonstrated intrarater reliability above a modest .54 on the strain scale. None of the raters had consistently poor intrarater reliability (i.e., $r < .50$) on all scales.

TABLE 1. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): Overall intrarater reliability coefficients for 21 raters.

Parameter	*r* (range)	No. of raters with *r* > .70
Overall severity	.57 (.21–.85)	2
Roughness	.77 (.49–.97)	14
Breathiness	.82 (.54–.99)	17
Strain	.35 (.16–.54)	0
Loudness	.78 (.34–.98)	7
Pitch	.64 (.55–1.00)	15

Table 2 shows the intrarater reliability coefficients (Spearman's rho) for the GRBAS scales, as well as the highest and lowest individual intrarater reliability coefficients. The strongest intrarater reliability pooled across raters was on the asthenia scale at .69, while the lowest was on strain at .53. Table 2 also reveals the number of raters whose intrarater reliability was greater than .70. For the GRBAS scale, eight raters achieved .70 or greater test–retest reliability on three scales: roughness, breathiness, and asthenia. Whereas on the CAPE-V task, two thirds of the raters achieved reliability of .70 or greater on three parameters, raters did not agree as consistently using the GRBAS, where only the breathiness parameter garnered .70 or greater reliability across just 50% of the raters.

Interrater Reliability

Interrater reliability was examined by calculating intraclass correlation (ICC) coefficients from two-way, random effects analysis of variance, where both raters and the vocal stimuli were treated as random sets from larger pools (Shrout & Fleiss, 1979; Winer, 1971). This calculation tests the level of agreement across all 21 raters. Table 3 shows ICC coefficients for the CAPE-V scales, with interrater reliability ranging from a high of .76 for overall severity to a low of .28 for pitch. Table 4 shows ICC coefficients calculated for interrater reliability on the GRBAS scale; the highest level was grade at .66, and the lowest was strain at .48. Between the CAPE-V and GRBAS instruments, four scales are most comparable: overall severity/grade, roughness, breathiness, and strain. Comparison of the values in Tables 3 and 4 reveals that among these four comparable scales, the interrater reliability values are slightly higher for the CAPE-V than for the GRBAS.

CAPE-V Versus GRBAS Judgments

The degree of association between the CAPE-V and GRBAS judgments was estimated by calculating the multiserial correlation (Harshbarger, 1977). The multiserial correlation estimates the association between one variable that is measured on an interval scale and another variable measured on an ordinal scale. The CAPE-V's 100-mm visual analog scale provides continuous, interval-level data (from 0 to 100), as measured directly from the line. In contrast, the GRBAS is an ordinal scale, requiring users to make rank-ordered judgments from 0 (*normal*) to 3 (*extreme*). The multiserial correlation was determined for each rater. Table 5 shows the average correlations and the range of individual values based on comparisons between four CAPE-V and GRBAS scales: overall severity/grade, roughness, breathiness, and strain. Overall severity and grade had the highest average correlation at .80. The two roughness scales were correlated at .76, the two breathiness scales at .78, and the two strain scales at .77.

TABLE 2. GRBAS (grade, roughness, breathiness, asthenia, strain): Overall intrarater reliability coefficients for 21 raters.

Parameter	r_s (range)	No. of raters with r_s > .70
Grade	.65 (.39–.87)	4
Roughness	.67 (.36–.92)	9
Breathiness	.67 (.22–.99)	11
Asthenia	.69 (.39–1.00)	8
Strain	.53 (.15–.75)	3

TABLE 3. CAPE-V interrater reliability.

Parameter	Shout–Fleiss ICC coefficients: Random set single rater
Overall severity	.76
Roughness	.62
Breathiness	.60
Strain	.56
Loudness	.54
Pitch	.28

Note. ICC = intraclass correlation.

Experienced Raters' Judgments Versus a Priori Severity Consensus Judgments

The overall correlation of the individual raters' judgments on grade with the a priori severity consensus judgments was .86. Table 6 shows the correspondence between the a priori consensus judgments of *normal, mild, moderate,* and *extreme* with the 21 raters' averages. This comparison (i.e., assessing the a priori consensus ratings in relation to our raters' judgments) can only be made with the GRBAS scale. To compare the a priori consensus ratings to the CAPE-V judgments, the CAPE-V continuous scale would have to be artificially subdivided into rank orders. The raters identified most but not all of the mild and moderately dysphonic voices in agreement with the a priori consensus judgments. As Table 6 indicates, only one voice in the mild category and one in the moderate category were not identified correctly with the consensus judgments; both of the single misidentified voices were judged by our raters to be in the lower, or less severe, group. However, voices judged a priori as normal and extreme were not categorized as well. More than two thirds of the normal voices (16 out of 22) were judged by our raters as mild, and about half (seven out of 13) of the extreme voices were identified as moderate.

Discussion

Auditory-perceptual scales are used clinically to assess voice quality. The CAPE-V, an auditory-perceptual voice

TABLE 4. GRBAS interrater reliability.

Parameter	Shout–Fleiss ICC coefficients: Random set single rater
Grade	.66
Roughness	.56
Breathiness	.59
Asthenia	.58
Strain	.48

TABLE 5. Average correlations between comparable CAPE-V and GRBAS scales.

CAPE-V	GRBAS	Multiserial correlation (range)
Overall	Grade	.80 (.52–.94)
Roughness	Roughness	.76 (.54–.92)
Breathiness	Breathiness	.78 (.61–.89)
Strain	Strain	.77 (.45–.91)

assessment instrument, was developed from a state-of-the-art understanding of the multidimensional factors that underlie psychophysical measurement and human perception (Kempster et al., 2009). To establish the empirical validity of the CAPE-V, we compared experienced raters' judgments of voice quality to judgments they made using another commonly used auditory-perceptual scale, the GRBAS. The extent to which raters' judgments can discern voice quality attributes reflects the instrument's content validity, while the agreement between raters' judgments using the CAPE-V versus GRBAS reflects empirical (concurrent) validity (Cronbach & Meehl, 1955). This study represents the first large effort to assess the reliability and empirical validity of the CAPE-V using experienced listeners from voice centers across the United States. This investigation is the largest such study to date, with 21 experienced raters, representing 17 separate facilities, judging 59 voice samples. These findings demonstrate that intra- and interrater reliability coefficients for the CAPE-V are slightly higher than those for the GRBAS. The strong multiserial correlations between the two scales suggest that the CAPE-V is empirically valid.

Because robust measurement reliability underlies any valid scale, we analyzed both intra- and interrater reliability among these 21 listeners. Our intrarater reliability correlations were lower than those reported by both Karnell et al. (2007) and Kelchner et al. (2010) for the CAPE-V. For the GRBAS, our intrarater reliability results were lower than those found by Karnell et al. (2007) but similar to those reported by De Bodt et al. (1997; see Tables 7 and 8). Our experienced raters found that the strain parameter for both the CAPE-V and GRBAS was the least perceptually salient dimension. The decreased intrarater reliability for these perceptual judgments corresponds to findings reported by both Kelchner et al. (2010) for the CAPE-V and De Bodt et al. (1997) for the GRBAS. Surprisingly, asthenia had the highest intrarater reliability value in this study, while it had the lowest in the De Bodt et al. (1997) study.

TABLE 6. Raters' grade judgments (GRBAS) versus a priori consensus severity judgments.

GRBAS grade judgments				A priori consensus severity judgments
Normal	Mild	Moderate	Extreme	
6	16	0	0	Normal = 22
1	**12**	0	0	Mild = 13
0	1	**10**	0	Moderate = 11
0	0	7	**6**	Extreme = 13
7	29	17	6	Total = 59

Note. Values in the diagonal (boldface) indicate how many of the raters' average judgments corresponded exactly to the a priori consensus severity judgments.

Our interrater reliability results reflect considerable variability across parameters, with the lowest correlation for strain. This variability is also demonstrated in both CAPE-V and GRBAS results reported by Kelchner et al. (2010) and De Bodt et al. (1997). In contrast, Karnell et al. (2007) reported consistently higher intra- and interrater reliability for judgments of grade (GRBAS) and overall severity (CAPE-V). However, that study employed only four raters from the same facility and calculated reliability using Spearman's rho.

Five factors may account for the differences in these reliability results compared to the previous studies. First, we employed a larger number of experienced raters than has been reported to date. Kreiman and Gerratt (2000) suggested that experienced listeners may introduce more variability into judgments of voice quality because they use a flexible strategy to determine salient perceptual features, making continual adjustments as they fine-tune their decisions. Consequently, this clinical experience may actually lower the reliability of judgments. Our experienced raters appeared to make consistent judgments regardless of scale. Second, while these raters were all voice specialists, their diverse background, training, and clinical settings may reflect inconsistencies that would be less likely among a cohesive group of raters working in the same clinic. Third, our raters made their judgments at two different time periods, using a standard protocol to ensure that both listening tasks were similar and unbiased by fatigue or order effects. Fourth, we averaged our data across raters, which may have obscured notable consistencies in individual perceptual decision making. Finally, we included 22 normal voices in our listening sample, to reflect the clinical spectrum of voice severity, which includes recovery to normal or near-normal vocal quality.

In addition to the primary goals, this study also provided an opportunity to consider possible alterations in the CAPE-V instrument. Originally, the conversational probe "Tell me about your voice problem" was selected for its expediency and ecological validity in a clinical voice evaluation. However, this task would not transfer appropriately to use with normal control speakers in a research study. An appropriate substitute might be to elicit spontaneous speech using a neutral question such as "Tell me about your favorite holiday" or "Describe the neighborhood where you grew up." Some participants identified potentially objectionable terms in two stimulus sentences. The /h/ onset–loaded phrase "How hard did he hit him?" presents an aggressive overtone. A suggested alternative might be "He helped Hannah hurry home." The glottal stop–loaded phrase "We eat eggs every Easter" includes a religious reference and could be modified to "We eat eggs every evening."

Did our study determine whether the CAPE-V is a valid instrument to assess auditory-perceptual voice quality? Our results do suggest that the CAPE-V is empirically valid when compared to the GRBAS. Although empirical validity does not mean that the CAPE-V is the same measure as the GRBAS, it does suggest that each tool measures similar

TABLE 7. Ranges of intra- and interrater reliability for the GRBAS scale across studies.

Data reported	De Bodt et al. (1997)	Karnell et al. (2007)	Present study
Listeners	ENTs and SLPs (*n* = 12)	SLPs (*n* = 4)	SLPs (*n* = 21)
Voices (*n*)	12	34	59
Intrarater reliability			
Statistic	κ	r_s	r_s
Minimum	.28 (asthenia)	.83 (grade)	.53 (strain)
Maximum	.70 (grade)	.91 (grade)	.69 (asthenia)
Interrater reliability			
Statistic	κ	r_s	ICC
Minimum	.17 (strain)	.80 (grade)	.48 (strain)
Maximum	.49 (grade)	.89 (grade)	.66 (grade)

Note. SLPs = speech-language pathologists.

constructs of vocal quality. However, a remaining obstacle is the known difficulty in determining phenotypes for disordered voice quality. Even though some attributes have physical correlates (e.g., pitch and frequency, loudness and intensity), the essential auditory-perceptual phenotypes for voice quality, such as roughness, breathiness, and strain, are multidimensional and covarying, and cannot be measured directly. Therefore, while this study establishes the empirical (concurrent) validity of the CAPE-V in relation to the GRBAS, there is no single criterion that can be used to establish construct validity.

Our literature is replete with attempts to use objective acoustic measures to cross-validate subjective auditory-perceptual judgments of voice quality, albeit with varying success. Some progress in this difficult task has emerged from an analysis-by-synthesis routine that manipulates acoustic signals to match perceived voice quality, with a goal of determining a consistent relationship (Bangayan, Long, Alwan, Kreiman, & Gerratt, 1997; Kreiman & Gerratt, 1996). Thus, another strategy for establishing the construct validity of the CAPE-V would be to evaluate predictive validity by comparing acoustic measures to raters' auditory-perceptual judgments.

Choosing an appropriate auditory-perceptual instrument to measure voice quality may also reflect the user's interests and capabilities. Clinicians and researchers have different goals and may subsequently prefer one scale over another, based on procedural or technical strengths and limits. It is critical that users define the elicitation protocol clearly (as needed for the GRBAS) or report any deviations from the published methods (as defined for the CAPE-V). Strengths of the CAPE-V include its use of a defined elicitation protocol, use of a consistent and ecologically valid conversational speech probe, and the inclusion of phonetically diverse speech contexts. In addition, the CAPE-V can be used to measure both prothetic and metathetic continua (Eadie & Baylor, 2006; Kent, 1996; Stevens, 1975). The CAPE-V's visual analog scales also yield interval-level data, allowing the use of parametric statistics, and thereby bringing greater power. Consequently, the CAPE-V may be more sensitive to small differences within and across auditory-perceptual judgments than the GRBAS would provide (Karnell et al., 2007). However, the GRBAS may be faster to administer in clinical settings. Globally, the GRBAS appears to be the most widely used auditory-perceptual tool.

Several limitations in this study warrant discussion. Although all of our samples were of adult voices, we were unable to match normal to abnormal voices by gender and age. Only a limited number of voice stimuli (*n* = 11) were repeated to assess intrarater reliability. Also, all of the raters in this study were deemed *experienced,* so we cannot predict how less experienced speech-language pathologists might judge voices using the CAPE-V scales. These limitations reduce the external validity of our findings. In future research, it would be helpful to gather clinical data to determine whether the CAPE-V can accurately capture incremental changes in voice quality across time. An optimal measure of voice quality would be able to document changes at the core of the

TABLE 8. Ranges of intra- and interrater reliability for the CAPE-V across studies.

Data reported	Kelchner et al. (2010)	Karnell et al. (2007)	Present study
Listeners	SLPs (*n* = 3)	SLPs (*n* = 4)	SLPs (*n* = 21)
Voices (*n*)	50 (pediatric)	34	59
Intrarater reliability			
Statistic	ICC%	r_s	*r*
Minimum	62% (strain)	.88 (severity)	.35 (strain)
Maximum	88% (breathiness)	.91 (severity)	.82 (breathiness)
Interrater reliability			
Statistic	ICC%	r_s	ICC
Minimum	35% (strain)	.86 (severity)	.28 (pitch)
Maximum	71% (breathiness)	.93 (severity)	.76 (severity)

clinical process, such as progress in behavioral voice therapy, fluctuations in laryngeal health, and gains in functional voice. This is an important consideration because the CAPE-V procedures specify that repeated stimuli obtained from the voice of a patient be directly compared from one assessment point to the next (see Kempster et al., 2009).

Sixteen out of 22 of the voices rated as normal by the listener were rated as mild by our 21 raters. There are several possible explanations for this surprising outcome. Because our raters knew that the listening samples included disordered voices, it is possible that they experienced some expectation bias toward greater vocal severity. Also, momentary irregularities in connected speech quality, such as intermittent glottal fry or roughness, may have influenced ratings of mild rather than normal. Finally, we know that our raters tended to avoid the endpoints of the perceptual scale, because they also judged about half (seven of 13) of the extreme voices as moderate.

A major issue that is likely to motivate future studies on the CAPE-V instrument itself is whether there are ways to improve the reliability and continue to establish the construct validity of this assessment procedure. Questions to test in future work include the following: Would CAPE-V reliability improve with training of listeners on the procedures and use of the scales? Would anchor stimuli in a training protocol be beneficial in improving reliability (Awan & Lawson, 2009; Chan & Yiu, 2006; Eadie & Baylor, 2006)? Is it helpful to clinicians to have the CAPE-V include the three speech contexts (vowels, sentences, and conversation) for measurement (de Krom, 1994; Zraick, Wendel, & Smith-Olinde, 2005)? Finally, are the scale values obtained using the CAPE-V clinically meaningful (Sechrest, 2005)?

The CAPE-V was developed to promote a standardized approach to evaluate and document auditory-perceptual judgment of voice quality (Kempster et al., 2009). This study provides evidence of its empirical validity, which justifies the use of the CAPE-V in clinical practice, educational programs, and professional development activities. Nonetheless, in future study, foremost is the need to further define the validity of CAPE-V in assessing this overlying construct of vocal quality. Thus, the quest to substantiate this methodology is a long-term process as the CAPE-V continues to be evaluated.

Acknowledgments

This research was supported by a noncompetitive project grant to the first author from ASHA's Special Interest Division 3 (Voice and Voice Disorders). The following individuals served in a volunteer capacity as expert raters in this study (listed in alphabetical order): Gayle Belin, Julie Blair, Linda Carroll, Kathleen Cox, Marina Gilman, Michael Hammer, Edie Hapner, Barbara Jacobsen, Christine Martin, Beth McWeeny, Diana Orbelo, Rita Patel, Brian Petty, Hylan Pickett, Julia Rademacher, Cara Sauder, Martin Spencer, Tonya Thomason, Michael Trudeau, Miriam van Mersbergen, and Nathan Welham. UAMS graduate students Jennipher Taylor, Sarah Kennedy, and Cherin Pace assisted with preparation of the study materials and data entry and management. Steve Gorman, Wendy DeLeo LeBorgne, and Jennifer Rettig at the Blaine Block Institute for Voice Analysis and Rehabilitation retrieved the dysphonic voice samples used in this study. Edward Carney, Research Associate at the University of Minnesota, Minneapolis, performed Harshberger's multiserial correlation to compare select data from CAPE-V and GRBAS judgments.

References

Awan, S., & Lawson, L. (2009). The effect of anchor modality on the reliability of vocal severity ratings. *Journal of Voice, 23,* 341–352.

Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., & Gerratt, B. (1997). Analysis bysynthesis of pathological voices using the Klatt synthesizer. *Speech Communication, 22,* 343–368.

Behrman, A. (2005). Common practices of voice therapists in the evaluation of patients. *Journal of Voice, 19,* 454–469.

Carding, P. N., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: State of the science review. *Journal of Laryngology and Otology, 123,* 823–829.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. In M. S. Lewis-Beck (Ed.), *Quantitative applications in the social science* (pp. 9–27). Thousand Oaks, CA: Sage.

Chan, K., & Yiu, E. (2006). A comparison of two perceptual voice evaluation programs for naïve listeners. *Journal of Voice, 20,* 229–241.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine, 119,* 166.e7–166.e16.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Croux, C. (1997). Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11,* 74–80.

Dejonckere, P. H., Obbens, C., de Moor, G. M., & Wieneke, G. H. (1993). Perceptual evaluation of dysphonia: Reliability and relevance. *Folia Phoniatrica et Logopaedica, 45,* 76–83.

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37,* 985–1000.

Delyiski, D. D., Shaw, H. S., & Evans, M. K. (2005). Influence of sampling rate on accuracy and reliability of acoustic voice analysis. *Logopedics, Phoniatrics and Vocology, 30*(2), 55–62.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., . . . Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39,* 155–164.

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527–544.

Gerratt, B., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America, 110,* 2560–2566.

Gorodetsky, R., Amir, G., & Yarom, R. (1992). Effect of ionizing radiation on neuromuscular junctions in mouse tongues. *International Journal of Radiation Biology, 61,* 539–544.

Harshbarger, T. R. (1977). *Introductory statistics: A decision map*. New York, NY: Macmillan.

Hirano, M. (1981). *Clinical examination of voice*. New York, NY: Springer-Verlag.

Karnell, M., Melton, S., Childes, J., Coleman, T., Dailey, S., & Hoffman, H. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21,* 576–590.

Kelchner, L. N., Brehm, S. B., Weinrich, B., Middendorf, J., deAlarcon, A., Levin, L., & Elluru, R. (2010). Perceptual evaluation of severe pediatric voice disorders: Rater reliability using the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice, 24,* 441–449. doi:10.1016/j.jvoice.2008.09.004.

Kelly, P. A., O'Malley, K. J., Kallen, M. A., & Ford, M. E. (2005). Integrating validity theory with use of measurement instruments in clinical settings. *Health Services Research, 40,* 1605–1619.

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kramer, J., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18,* 124–132.

Kent, R. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23.

Kreiman, J., & Gerratt, B. (1996). Perceptual structure of pathologic voice quality. *The Journal of the Acoustical Society of America, 100,* 1787–1795.

Kreiman, J., & Gerratt, B. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104,* 1598–1608.

Kreiman, J., & Gerratt, B. (2000). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America, 108,* 1867–1876.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122,* 2354–2364.

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,* 21–40.

Lazarus, C. L. (2009). Effects of chemoradiotherapy on voice and swallowing. *Current Opinion in Otolaryngology & Head and Neck Surgery, 17,* 172–178.

Marks, L., & Algom, D. (1998). Psychophysical scaling. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 81–178). San Diego, CA: Academic Press.

Sechrest, L. (2005). Validity of measures is no simple matter. *Health Services Research, 40,* 1584–1604.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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,* 323–335.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Stevens, S. S. (1975). *Psychophysics*. New York, NY: Wiley.

Verdolini, K., Rosen, C. A., & Branski, R. C. (2006). *Classification manual for voice disorders—I*. Mahwah, NJ: Erlbaum.

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Otorhinolaryngology, 261,* 429–434.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York, NY: McGraw-Hill.

Zraick, R. I., Wendel, K. W., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574–581.

Received December 3, 2009
Revision received May 3, 2010
Accepted July 28, 2010
DOI: 10.1044/1058-0360(2010/09-0105)

Contact author: Richard I. Zraick, University of Arkansas for Medical Sciences, Mail Slot 772, 4301 West Markham Street, Little Rock, AR 72205. E-mail: zraickrichardi@uams.edu.