The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) Psychometric Characteristics: II European Portuguese Version (II EP CAPE-V)

*Sancha C. de Almeida, †Ana P. Mendes, and ‡Gail B. Kempster, *Lisbon and †Setúbal, Portugal, and ‡Chicago, Illinois

Summary: Summary objective: The purpose of this study was to assure a reliable and valid European Portuguese (EP) version of Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). This resulted in the second EP version of CAPE-V (II EP CAPE-V), with permission granted by American Speech-Language-Hearing Association.
Study design: This is a transversal, observational, descriptive, and comparative study.
Methods: Retranslation of CAPE-V into EP was reviewed by an EP linguistic expert for content validity. A total of 20 subjects: 10 male individuals (mean age = 45) and 10 female individuals (mean age = 43) formed a control group (n = 10) and a dysphonic group (n = 10) were matched by age and gender. All subjects' CAPE-V phonatory tasks were captured with PEYLE PMENI (China) microphone and recorded with TASCAM DR-05 (Tokyo, Japan). Fourteen speech-language pathologists voice experts (>5 years of clinical practice) rated 26 voice samples with 6 repeated samples added for intrarater reliability. All voice samples were heard using AKG K101 (Europe) headphones and were rated in two sessions with a 1-week interval: one with the II EP CAPE-V; and the second with the GRBAS scale to establish for inter-rater reliability and construct and concurrent validity. Statistical analysis for inter-rater reliability was obtained with the intraclass correlation coefficient. Intrarater reliability was obtained with Pearson correlation. Construct and concurrent validity were performed with Student *t* test and multiserial correlation coefficient, respectively. *SPSS* 22.0 (IBM Corp, Armonk, NY) and *LISREL* 8.8 (Scientific Software International, Inc, Chicago, IL) were used with significance level cut-off points: $r > 0.70$ and $\alpha = 0.05$.
Results: High inter-rater reliability was obtained for all vocal parameters (intraclass correlation coefficient > 0.84) revealing good equivalence. Intrarater reliability was high ($r > 0.87$) for overall severity, breathiness, and pitch; good ($r = 0.73$) for strain; and moderate ($r > 0.61$) for roughness and loudness. These results revealed good reproducibility and stability of the II EP CAPE-V over time. Content validity was assured by an EP linguistic expert. Construct validity was obtained for all vocal parameters ($P < 0.05$), except for strain ($P = 0.52$), revealing these were the salient parameters for rating normal and dysphonic voices samples. Concurrent validity between the II EP CAPE-V and the GRBAS scales had strong correlations ($r > 0.89$) for overall severity/grade, roughness, and breathiness parameters, suggesting both instruments measure the same construct.
Conclusions: The II EP CAPE-V is a reliable and valid instrument for auditory-perceptual evaluation of the EP population, with all psychometric measures assured.
Key Words: CAPE-V−Voice evaluation−Auditory-perceptual evaluation−Voice disorders−Voice quality.

INTRODUCTION

Auditory-perceptual evaluation is part of a multidimensional voice evaluation.[1] This is a worldwide valued procedure,[1−4] and it is claimed to be the "golden standard" to document voice disorders.[3,5] In an effort to standardize auditory-perceptual evaluation, different schemes and scales have been specifically designed. The grade, roughness, breathiness, asthenia, strain (GRBAS) scale[6] and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)[7] are widely used by health and/or educational professionals in the field of voice (ie, speech-language pathologists [SLPs], otorhinolaryngologist (ENT), and voice teachers). The scale used is selected depending on specific clinical or research purposes.[8]

According to the current standards of evidence-based medicine, any health status evaluation instrument must be reliable and valid to be clinically useful.[1,9] These psychometric characteristics assure the integrity and quality of a measurement instrument.[10,11] Reliability is the degree to which an instrument is free from random error or the extent to which obtained scores can be reproduced.[9,12,13] Inter-rater reliability determines the equivalence of ratings obtained with an instrument when used by different raters.[11] Intrarater reliability or test-retest reliability is the reproducibility or stability measure of an instrument over time.[9,13] This reliability is determined by the administration of the same instrument to the same group of raters at two different moments.[9−11,13] For both inter- and intrarater reliability, the acceptable coefficient threshold levels are 0.70 for group comparisons and 0.90-0.95 for individual measurements over time.[9,13] Validity of an instrument is the degree to which the instrument measures what it purports to measure.[9,11−13] The Scientific Advisory Committee of the Medical Outcomes Trust determined that validity has three

Accepted for publication February 14, 2018.
From the *Hospital da Luz—ENT Department, Lisbon, Portugal; †Health Science School of Polytechnic Institute of Setubal, Setúbal, Portugal; and the ‡Department of Communication Disorders and Sciences, Rush University Medical Center, Chicago, Illinois.
Address correspondence and reprint requests to Sancha C. de Almeida, Serviço de Otorrinolaringologia, Hospital da Luz, Avenida Lusíada, 100, 1500-650 Lisbon, Portugal. E-mail: scalmeida@hospitaldaluz.pt
Journal of Voice, Vol. 33, No. 4, pp. 582.e5−582.e13
0892-1997

https://doi.org/10.1016/j.jvoice.2018.02.013

TABLE 1.
CAPE-V Inter-rater Reliability Across Different Studies

Study	Statistical Analysis (>0.70)	Vocal Parameters					
		Overall Severity	Roughness	Breathiness	Strain	Pitch	Loudness
Karnell et al (2007)	*r*	0.88	NA	NA	NA	NA	NA
Jesus et al (2009a)	*P*	0.964	0.834	0.991	0.659	0.500	1.000
Kelchner et al (2010)	ICC	67%	68%	71%	35%	68%	63%
Zraick et al (2011)	ICC	0.76	0.62	0.60	0.56	0.54	0.28
Nemr et al (2012)	ICC	0.911	0.870	0.897	0.828	NA	NA
Mozzanica et al (2013)	ICC	0.92	0.91	0.90	0.76	0.83	0.82
Núñez-Batalla et al (2015)	ICC	>0.833	>0.750	>0.769	>0.648	>0.710	>0.545

Abbreviation: NA, not available.

aspects: content, construct, and criterion.[9] Content validity reflects the items included adequacy to the domain of the instrument.[10] Construct validity is the degree to which an instrument measures the construct under study.[14] Concurrent validity is a type of criterion-related validity, where evidence is showed by the extent to which the scores of the instrument are related to a criterion measure.[9,13] To determine this validity, scores of an instrument are correlated to the scores of another one that measures the same construct on the same subjects.[11]

Auditory-perceptual evaluation is often considered to be subjective and influenced by several factors such as listeners (eg, listeners experience and background), voice stimuli, and rating scales.[3,4,15–25] However, variability can be minimized when influential known factors are identified and experimental procedures are well designed and controlled.[3,26]

The CAPE-V's psychometric characteristics have been well reported in the literature.[8,26–33] CAPE-V inter-reliability was high (>0.70) for overall severity, roughness, breathiness, strain, pitch, and loudness parameters (Table 1). High intrarater reliability (>0.70) was reported for CAPE-V parameters, across several studies (Table 2). CAPE-V content validity was assured by different professionals (eg, SLPs, linguistics, and phoniatrics), depending on the language translation (eg, Brazilian Portuguese, European Portuguese [EP], Italian, and Spanish). CAPE-V construct validity was assured for the Italian and Brazilian Portuguese CAPE-V versions.[8,30] Student *t* test results revealed significant differences between normal and dysphonic voice samples for all six CAPE-V parameters ($P < 0.0001$) for both Italian and Brazilian versions. Concurrent validity was found based on the high correlations ($r > 0.70$) between four comparable CAPE-V and GRBAS parameters: overall severity/grade, roughness, breathiness, and strain (Table 3).

Nevertheless, several methodological and statistical differences could have influenced CAPE-V psychometric measures previously reported. Inter-rater reliability was often based on a reduced number of listeners (<4),[8,26–30,32] which limited the power of results. Zraick et al[33] were the only authors who tested a large number of listeners (n = 21) for inter-rater reliability. Most CAPE-V studies used SLPs who specialized in voice disorders with more than 5 years of clinical experience.[8,29–31,33] Usually, experienced listeners reveal better inter-rater reliability when compared to inexperienced listeners.[24,25,34,35] Some studies provided only dysphonic voices samples to the listeners,[26,27,29] whereas others provided normal and dysphonic voice samples.[8,30,32,33] The study of Karnell et al[28] was the only investigation that provided balanced voice samples, matched by age and gender. Voice samples have been presented to listeners in the same sequence,[26,27,29,30,33] or in two different randomized sequences.[8,28,32] In most studies, repeated voice ratings were separated by a 1-week interval.[28–30,32]

Intrarater reliability procedures also varied across several studies. In some studies, listeners judged all voice samples twice,[28,30] whereas in other studies, they judged a subset of total voice sample.[8,29,32,33] The number of phonatory tasks

TABLE 2.
CAPE-V Intrarater Reliability Across Different Studies

Study	Statistical Analysis (>0.70)	Vocal Parameters					
		Overall Severity	Roughness	Breathiness	Strain	Pitch	Loudness
Karnell et al (2007)	*r*	>0.88	NA	NA	NA	NA	NA
Kelchner et al (2010)	ICC	87%	82%	82%	63%	78%	79%
Zraick et al (2011)	*r*	0.57	0.77	0.82	0.35	0.78	0.64
Nemr et al (2012)	ICC	0.927	NA	NA	NA	NA	NA
Mozzanica et al (2013)	ICC	0.92	0.92	0.90	0.89	0.88	0.80
Núñez-Batalla et al (2015)	ICC	>0.972	>0.969	>0.952	>0.921	>0.894	>0.851

Abbreviation: NA, not available.

TABLE 3.
CAPE-V and GRBAS Concurrent Validity Across Different Studies

Study	Statistical Analysis (>0.70)	Vocal Parameters			
		Overall Severity/Grade	Roughness	Breathiness	Strain
Karnell et al (2007)	*r*	0.95	0.90	0.89	0.91
Jesus et al (2009b)	ρ	0.60	0.26	0.80	NA
Zraick et al (2011)	*r*	0.80	0.76	0.78	0.77
Nemr et al (2012)	*r*	0.80	NA	NA	NA
Mozzanica et al (2013)	*r*	0.92	0.84	0.87	0.79
Núñez-Batalla et al (2015)	ICC	0.874	0.849	0.612	0.843

Abbreviation: NA, not available.

also varied. Some used all three CAPE-V phonatory tasks,[8,28,30–32] whereas others used only some of them: sustained vowels, reading aloud sentences and a text[26,27], repeating aloud sentences,[29] or spontaneous speech.[33]

The CAPE-V's content validity was analyzed into its translation and adaptation in different languages: Brazilian Portuguese, EP, Italian, and Spanish. This was ensured by different professionals, depending on the language translation. The first CAPE-V translation into EP was performed by Jesus et al.[27] However, the sentences developed did not accomplish the sentences' original purposes, with the phonetic targets specified in the original CAPE-V. Thus, this translation did not guarantee the content validity in comparison to the original instrument.

CAPE-V construct validity has little data available, since the construct validity of an instrument is often developed over time. The Italian and Brazilian Portuguese studies were the only ones that reported this CAPE-V psychometric measure.[8,30]

The present study aimed to develop a reliable and valid second EP version of the original CAPE-V,[7,36] incorporating sentences meeting the original CAPE-V's purposes and where all the psychometric measures were assured: inter- and intrarater reliability, as well as content, construct, and concurrent validity.

METHODS

This was a transversal, observational, descriptive, and comparative study. All subjects signed an informed consent approved by the Ethics Committee of Hospital da Luz (CES/006/2015/PA, 2015).

Speakers

Twenty speakers participated in this study: 10 male individuals (mean age = 45) and 10 female individuals (mean age = 43), divided into a control group (CG = 10) and a dysphonic group (DG = 10); subjects were matched by age and gender (Table 4). Inclusion criteria for the CG were (1) no organic or functional laryngeal disorder confirmed by indirect laryngoscopy; (2) native EP speaker; (3) over 18 years old; and (4) adequate literacy ability for the purposes of the study. Criteria for DG were (1) presence of organic or functional laryngeal disorder confirmed by indirect laryngoscopy; (2) native EP speaker; (3) over 18 years old; and (4) with adequate literacy ability for the purposes of the study. Exclusion criteria were (1) a history of cognitive, or speech and language disorders; and (2) allergy, vocal complaints, and/or breathing problems on the day of voice recording. The samples forming both groups were non-randomized, convenient samples.

TABLE 4.
Speakers CG and DG Matched by Age and Gender

Gender	Age (y)	CG	DG
M	34	1	1
	37	1	1
	42	1	1
	52	1	1
	61	1	1
F	30	1	1
	34	1	1
	44	1	1
	52	1	1
	55	1	1
Total		10	10

Abbreviations: F, female; M, male.

Listeners

Fourteen SLPs were recruited as listeners. Inclusion criteria were (1) more than 5 years of voice clinical experience; (2) weekly voice patient caseload; (3) bilateral normal hearing limits for speech purposes; (4) experience using the CAPE-V instrument for the evaluation of voice; (5) experience using the GRBAS scale; and (6) native EP speaker. Exclusion criteria were (1) history of cognitive or speech and language disorder. Two men (mean age = 28) and 12 women (mean age = 38) participated as listeners. The average length of their clinical voice experience was 11 years (Table 5).

Equipment

Voice samples were captured with a PYLE PMEMI (China) headset microphone, electret condenser, omnidirectional with frequency response 20 Hz-20 KHz and sensitivity −44 dB ± 3 dB, and recorded on a TASCAM DR-05 (Tokyo, Japan) portable digital recorder, 16 bits, mono, with a sample frequency of 44,100 Hz. Ambient noise was always below 50 dB, measured by a Rolls SLM305 digital sound level meter.

TABLE 5.
Distribution of Listener's Subjects by Age and Year of Experience

Age (y)	Gender	n	Experience (y)	Gender	n
17-29	M	2	3–5	M	0
	F	4		F	2
30-39	M	0	6–10	M	2
	F	4		F	5
40-49	M	0	10–20	M	0
	F	1		F	2
50-59	M	0	>20	M	0
	F	2		F	3
>60	M	0			14
	F	1			
Total		14			14

Abbreviations: F, female; M, male.

Instruments

For this study two instruments were used to collect audio-perceptual data: II EP CAPE-V (Appendix A) and GRBAS.[6] II EP CAPE-V is an instrument with determined voice data collection and scoring procedures. Voice sample was composed by three phonatory tasks: sustained [a, i] three times each for 3-5 seconds, reading aloud six sentences, and 20 seconds of spontaneous speech in response to the prompt "Tell me about the place where you grew up." Listeners judged the following six vocal parameters: (1) overall severity, (2) roughness, (3) breathiness, (4) strain, (5) pitch, and (6) loudness using a visual-analog scale of 100 mm. Deviance degree was marked using a tick mark on the scale. Leftmost portion of the line reflected normal voice or nonexistence of a vocal parameter. The right end of the scale reflected the most extreme deviance perception. Below the visual-analog scale line, general regions were evenly displayed as supplement categorical severity indicator (eg, "MI" mildly deviant, "MO" moderately deviant, and "SE" severely deviant), using terminology familiar to listeners.

The GRBAS scale[6] does not have a specific procedure protocol for voice sample collection, documentation, or evaluation. This scale evaluates the following audio-perceptual vocal parameters: grade (G), rough (R), breathy (B), asthenic (A), and strained (S), in a four-point, ordinal scale from 0 (normal) to 3 (extreme).

Procedures

The flowchart of Figure 1 represents the research design of II EP CAPE-V and the sequence of procedures followed during this study.

CAPE-V retranslation

A psychometric analysis of the first EP CAPE-V translation[27] revealed that the sentences proposed mirror neither all the CAPE-V original sentences' purposes nor the phonetic targets. A permission to retranslate and adapt the CAPE-V into EP

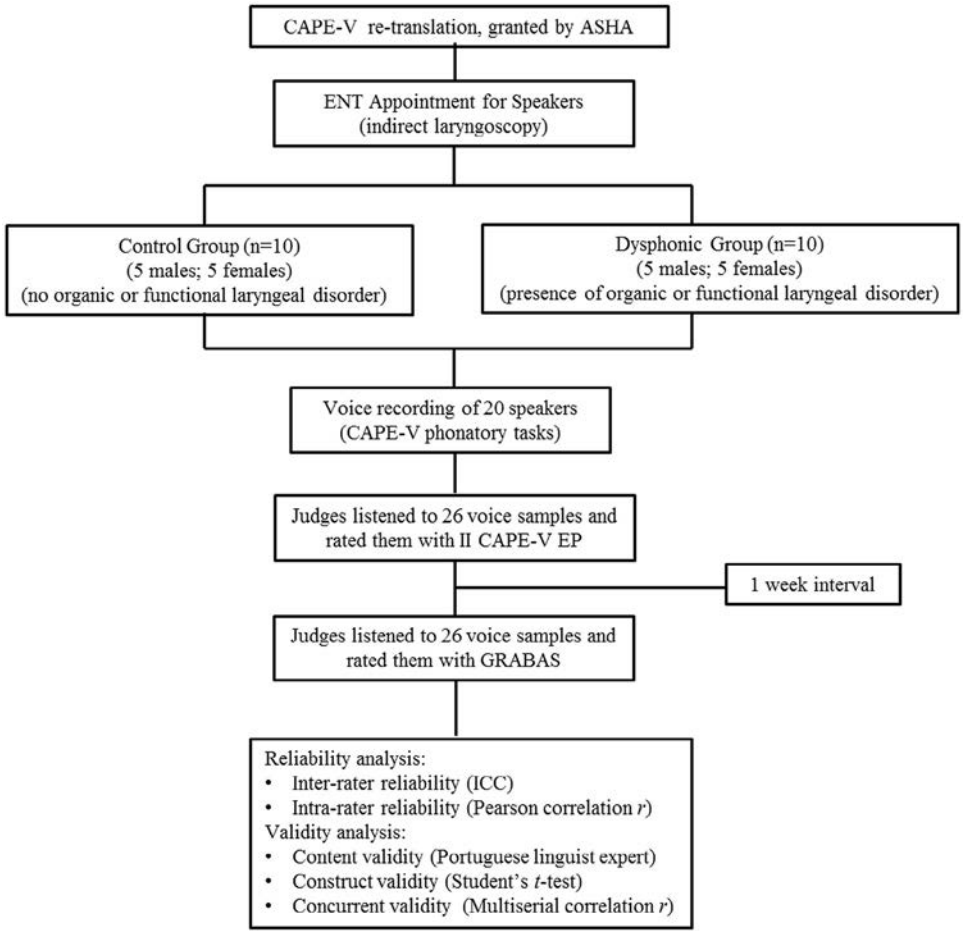

FIGURE 1. II EP CAPE-V research design and sequence procedures.

was granted by American Speech-Language-Hearing Association. II CAPE-V was designed to fulfill the requirements established under the original CAPE-V.[7,36] The six new sentences proposed for the reading aloud task were conceptualized and adapted to EP linguist context and reviewed by a Portuguese linguist. The following were the new sentences:

Sentence (A) *Num domingo esteve sol e fui com o avô António à explanada Évora comer uma empada* (On Sunday, it was sunny and I went with grandfather António to the terrace of the *Évora* cafe to eat a pie). This sentence target was to examine the coarticulatory production of all oral and nasal EP vowels.
Sentence (B) *Segundo Simão, só Samuel sabe* (According to Simão, only Samuel knows). This sentence target was to assess soft glottal attacks and voiceless to voiced transitions through a sentence that only contains words that emphasize easy onset with (s).
Sentence (C) *A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa* (Zé, Gabriel's mother, gave him an orange cake and old wine from *Runa*). This sentence target was to assess possible voiced stoppages/spasms through the production of all EP voiced phonemes.
Sentence (D) *É hora da Urraca ir à caça* (It is time for Urraca to go hunting). This sentence target was to assess possible hard glottal attacks through words beginning with vowels that may elicit a hard glottal attack.
Sentence (E) *Onde eu brinco há um ninho de andorinhas encostado ao muro* (Where I play, there is a swallow's nest next to the wall). This sentence target was to assess hyponasality and possible stimulability for resonant voice therapy through the production of all the EP nasal vowels and consonants.
Sentence (F) *A Kika tapou a tua capa preta* (Kika covered your black cape). This sentence target was to assess intraoral pressure and possible hypernasality or nasal air emission through the production of voiceless plosive sounds.

For spontaneous speech elicitation, the quote *Fale-me do sítio onde cresceu* (Tell me about the place where you grew up) was used. This was similar to the CAPE-V standardized procedures.[33]

Voice recording

All phonatory tasks were recorded following CAPE-V instructions.[7,36] The same recordings and tasks procedures were used to obtain all the voice samples from all the speakers. Equipment was always tested and calibrated with a 500 Hz reference pure tone, confirmed by acoustic analysis. This tone was recorded and analyzed at the beginning of each recording day.

Speakers were seated in a comfortable position. Voice samples were recorded in a sound treated room at the ENT Department at Hospital da Lu. Ambient noise was always below 50 dB[37] as measured with a Rolls SLM305. Voice productions were recorded directly on a TASCAM DR-05. A PEYLE PMENI headset microphone was positioned at a constant distance of 4 cm from speaker's mouth on a 45° angle from the mouth.[37] Similar to the study of Zraick et al,[33] voice samples were not normalized for intensity and noise reduction. After each recording, voice samples were labeled with no speaker identification information.

Listening

Twenty voice samples (10 normal and 10 dysphonic) were present once to each listener. Six of them were repeated and randomly mixed with the previous ones to enable test-retest to determine the intrarater reliability.

Before the first listening session, all the listeners underwent a pure tone hearing screening at 20 dB HL at 500, 1000, 2000, and 4000 Hz,[38] to assure that hearing was within normal limits. During the first listening session, 14 listeners rated 26 voice samples using the II EP CAPE-V on paper (Appendix A). One week later, they rated the same voice sequence using the GRBAS scale[8,30] on paper. Voice samples were presented in a quiet room with ambient noise <50 dB, at the ENT Department at Hospital da Luz. Each listener was comfortably seated at a computer, equipped with AKG K101 (Europe) headphones[8,29,39] and was allowed to adjust the volume to a comfortable listening level.[33] Each listener was allowed to listen to the voice samples more than once[8,33] but no more than three times. Voices were reproduced in four blocks of (1) seven voice samples, (2) six voice samples, (3) seven voice samples, and (4) six voice samples, with a 10 minutes interval between each block[8] to reduce fatigue and inattentiveness.

Statistical analysis

Inter-rater reliability was calculated with intraclass correlation coefficient (ICC) following a two-way mixed effects model.[40] A confidence interval of 95% was used. Intrarater reliability was performed with Pearson correlation coefficients ($r > 0.70$) for all vocal parameters. For reliability analyses, all calculations were performed on *SPSS* 22.0 statistical software (IBM Corp, Armonk, NY).

Construct validity was based on a contrasted groups approach. The independent-samples Student *t* test was used to compare means between the two groups, across all the vocal parameters. This analysis was performed using *SPSS* 22.0.

Concurrent validity was estimated with a multiserial correlation coefficient for each listener and for average scores of the total listeners with $r > 0.70$. This correlation estimated the association degree between a CAPE-V interval variable and a GRBAS ordinal variable.[41] This analysis was performed on *LISREL* 8.80 software (Scientific Software International Inc., Chicago, IL).

RESULTS

Inter-rater reliability was obtained using ICC for each CAPE-V vocal parameter. There was a high level of agreement (ICC > 0.84) across all 14 listeners for all the vocal parameters (Table 6). Overall severity presented the highest ICC (ICC = 0.96) and strain the lowest (ICC = 0.84).

Intrarater reliability was determined for all vocal parameters. Average, highest, and lowest individual of the intrarater reliability coefficients (Pearson *r*) were calculated (Table 7). Overall severity, breathiness, and pitch parameters revealed high intrarater reliability ($r > 0.87$), whereas strain ($r = 0.73$) was good, and roughness and loudness reflected only moderate intrarater reliability ($r = 0.61$ and $r = 0.69$, respectively). Seven of the 14 listeners revealed good intrarater reliability (>0.70) for overall severity, breathiness, and loudness parameters.

TABLE 6.
Inter-rater Reliability of II EP CAPE-V Parameters for 14 SLPs Listeners

Vocal Parameter	ICC
Overall severity	0.96
Roughness	0.92
Breathiness	0.95
Strain	0.84
Pitch	0.86
Loudness	0.90

TABLE 7.
Intrarater Reliability of II EP CAPE-V Parameters for 14 SLPs Listeners

Vocal Parameters	*r*
Overall severity	0.87
Roughness	0.61
Breathiness	0.87
Strain	0.73
Pitch	0.92
Loudness	0.69

Construct validity was estimated by comparing CG and DG mean scores and standard deviations using independent-sample Student *t* test, for all CAPE-V parameters (Table 8). For all vocal parameters, mean scores and standard deviations of the DG were higher than for the CG. There were significant differences between DG and CG ($P < 0.05$) for overall severity, roughness, breathiness, loudness, and pitch. Strain was the only parameter that did not reach significance between groups ($P = 0.52$).

Concurrent validity

Concurrent validity was obtained by a multiserial correlation between the four comparable II EP CAPE-V, and GRBAS parameters were determined for each listener, as well as for the average scores of the total of listeners. Three parameters presented strong and positive correlations ($r > 0.89$). They were (1) overall severity/grade, (2) roughness, and (3) breathiness. Strain did not meet the threshold determined by the investigators ($r < 0.70$) (Table 9).

TABLE 8.
CG and DG Means and Standard Deviations of II EP CAPE-V Parameters

Vocal parameter	Control Group Mean ± SD	Dysphonic Group Mean ± SD	*P*-value
Overall severity	12.77 ± 11.88	38.24 ± 21.04	0.01*
Roughness	13.68 ± 7.92	39.01 ± 11.49	0.00*
Breathiness	12.77 ± 11.88	38.24 ± 21.04	0.01*
Strain	23.04 ± 12.87	26.59 ± 11.06	0.52
Pitch	7.98 ± 5.18	20.29 ± 10.41	0.01*
Loudness	9.62 ± 5.59	20.26 ± 13.59	0.04*

* Significant differences.
$P < 0.05$.
Abbreviation: SD, standard deviation.

TABLE 9.
Multiserial Correlation Between II EP CAPE-V and GRBAS Parameters

CAPE-V	GRBAS	Multiserial correlation
Overall severity	Grade	0.95
Roughness	Roughness	0.89
Breathiness	Breathiness	0.90
Strain	Strain	0.47

DISCUSSION

Auditory-perceptual evaluation plays an essential role in multidimensional voice evaluation[1] and in establishment of a voice therapy plan.[2] Different auditory-perceptual instruments may be selected depending on specific clinical or research purposes. The CAPE-V[7] is a more recent instrument compared to the well-known GRBAS scale.[6] Its usage has been increasing in both clinical and research settings. CAPE-V psychometric characteristics are well reported as recommended by Scientific Advisory Committee of the Medical Outcomes Trust.[8,26–33]

Reliability is a necessary psychometric measure as it establishes the degree in which an instrument is free from random error and the extent to which results can be reproduced. II EP CAPE-V high inter-rater reliability (ICC > 0.84) was obtained for all parameters, revealing strong agreement among 14 listeners. The present study, revealed the highest ICC of agreement for all the vocal parameters when compared to other studies (Table 1). These results may be due to the number of listeners (n = 14) used when compared to the most common in CAPE-V studies (n < 4 listeners).[8,27–30,32] Zraick et al[33] presented the largest number of listeners (n = 21). Nevertheless, the number of EP voice experts in Portugal is lower than in United States, thus this factor does not diminish inter-reliability value. In the current study, the 14 listeners were SLPs voice experts, with more than 5 years of clinical practice, reflecting the high inter-rater reliability common for experienced listeners.[24,33–35] Voice stimuli included the three CAPE-V phonatory tasks and were produced by 10 normal and 10 dysphonic subjects, matched for age and gender. These balanced voice stimuli may have contributed to low variability across the CG and DG, resulting in a better inter-rater agreement. II EP CAPE-V inter-rater reliability results demonstrate that 14 EP expert listeners rated CAPE-V vocal parameters consistently, independently of their different backgrounds, clinical settings, and internal standards.

High intrarater reliability was obtained on overall severity, breathiness, and pitch parameters ($r > 0.87$). Good reliability was found on strain ($r = 0.73$), and moderate intrarater reliability on roughness and loudness ($r > 0.61$). These findings revealed the stability of each listener's rating for those vocal

parameters. These results could have been influenced by at least two factors: (1) the number of repeated voice samples and (2) the rating procedures used. In the present study, 30% of voice samples were rerated by each listener increasing intrarater reliability representativeness. All voice samples were rated in two sessions with a 1-week interval. On the first session, all voice samples were rated with II EP CAPE-V, whereas on the second with GRBAS. This guaranteed that listeners underwent the same conditions, therefore, minimizing the possibility of internal standards changing over time. Pitch intrarater reliability was higher ($r = 0.92$) than the reported in the literature.[8,29,30,32,33] This result showed that pitch was a remarkable and stable auditory-perceptual parameter for EP listeners. Intrarater variability could be influenced by a listener's experience.[24,35] However, these 14 listeners were voice experts with at least 5 years of clinical experience, similar to most of CAPE-V intrarater reliability studies.[8,29,30,33] This was the first CAPE-V intrarater reliability study applied to EP voice samples by EP listeners. Results indicated that EP listeners displayed stable internal standards, demonstrating their intrarater reliability for auditory-perceptual evaluation.

II EP CAPE-V content validity was assured by the review and adaptation of reading aloud and spontaneous speech tasks. For the II EP CAPE-V version, all sentences were reviewed by an EP linguistic expert who assured that they fulfilled the same purposes and phonetic targets established by the original CAPE-V,[7,36] and in line with EP linguistic and cultural context. This psychometric measure assured that II EP CAPE-V followed the same procedures and targets, regardless of the language into which CAPE-V was translated.

For construct validity of II EP CAPE-V, statistically significant differences were found between CG and DG for overall severity, roughness, breathiness, pitch, and loudness parameters ($P < 0.05$). These results were similar to Mozzanica et al[30] and Nemr et al[31] The strain parameter also revealed differences between the CG and the DG with higher mean scores for DG (mean = 26.59) than for CG (mean = 23.04). These mean differences contributed to the identification of a voice disorder for a given speaker. Strain had the highest mean scores (ranged from 7.98 to 23.04) and standard deviations (ranged from 5.18 to 12.87) in the CG, similar to those reported by Nemr et al.[31] These results suggest that strain is not a valuable auditory-perceptual parameter to differentiate normal from dysphonic voices of EP population. This present study is innovative and relevant for both national and international clinical and research endeavors because it contributed to establishing CAPE-V construct validity, where little data are available.

II EP CAPE-V concurrent validity was established based on multiserial correlations between the four comparable II CAPE-V and GRBAS parameters. Listeners rated voice samples with II EP CAPE-V and 1 week later with GRBAS, avoiding potential crossover effect. Results revealed high correlations between overall severity/grade, roughness, and breathiness ($r > 0.89$). These results were similar to those reported by Karnell et al[28] and higher than those reported in other studies[26,30,32,33] (Table 3). II EP CAPE-V and GRBAS strain correlation was lower ($r = 0.47$) when compared to the previous studies.[28,30,32,33] This was in agreement with previous construct validity results where no significant differences were found between the CG and the DG. Results reported in this study provide evidence that the CAPE-V and the GRBAS scales measure similar constructs, contributing to the establishment of II EP CAPE-V concurrent validity. These results support the use of the CAPE-V for auditory-perceptual voice evaluation and voice therapy outcomes measurement in national and international studies. The CAPE-V has formal administration procedures with specific phonatory tasks, encouraging clinicians to follow a standard auditory-perceptual voice evaluation protocol. When selecting an instrument, the examiner must consider the psychometric characteristics as well as the advantages and disadvantages of each, depending on the purpose of the assessment.

Limitations of this study can be related to methodological procedures. A smaller number of voice samples (n = 20) was used in comparison to other CAPE-V studies.[8,26,28–30,32,33] Nevertheless, speakers were selected according to laryngoscopic results and were matched by age and gender. All listeners were experts in voice disorders with an average of 11 years of clinical practice. No anchor stimuli were provided before II EP CAPE-V rating session to calibrate unacquainted listeners. This may had some impact on reliability and validity, although the original CAPE-V rationale and protocol procedures[7,36] were closely followed.

For future research, the impact of SLPs experience in psychometric measures (ie, inter- and intrarater reliability and construct validity) could be studied to determine the need of auditory-perceptual training courses, as well as number of training hours. Further studies should include together both auditory and visual stimuli (eg, seeing the patient in person), which may improve the understanding of the dimension of strain and how it is rated. Lastly, the phonatory task individual sensitivity could be studied to determine which phonatory task is most efficient for the auditory-perceptual evaluation purposes.

CONCLUSIONS

The present study revealed that II EP CAPE-V is a reliable and valid instrument for auditory-perceptual voice evaluation of EP population. This study measured II EP CAPE-V inter- and intrarater reliability, as well as content, construct, and concurrent validity. The reported results underscore the national and international establishment of important psychometric characteristics of the CAPE-V, supporting its continued use in educational, clinical, and research fields.

Acknowledgments

We thank the SLPs who freely collaborated in this study. Thanks to Fernando Martins, PhD for the CAPE-V's translation into EP. We are grateful to Margarida Lemos, PhD and Maria Fátima Salgueiro, PhD for the statistical analyses.

APPENDIX

Appendix A. II EP CAPE-V

"Consensus Auditory-Perceptual Evaluation of Voice"
2ª Versão Português Europeu (II CAPE-V PE)

Legenda: C = consistente I = inconsistente
DL: Desvio ligeiro
DM: Desvio moderado
DS: Desvio severo

Pontuação

Grau de severidade global _____ C I ____/100
DL DM DS

Rouquidão _____ C I ____/100
DL DM DS

Soprosidade _____ C I ____/100
DL DM DS

Tensão _____ C I ____/100
DL DM DS

Altura tonal (indicar o tipo de alteração): grave/agudo

_____ C I ____/100
DL DM DS

Intensidade (indicar o tipo de alteração): fraca/forte

_____ _____ C I ____/100
DL DM DS

_____ _____ C I ____/100
DL DM DS

_____ _____ C I ____/100
DL DM DS

COMENTÁRIOS SOBRE A RESSONÂNCIA: Normal Alterada (breve descrição): _____
_____

FACTORES ADICIONAIS (por ex.: diplofonia, aspereza, falseto, astenia, afonia, bitonalidade, tremor, estridência, "glottal fry", outros aspectos relevantes) _____
_____

II CAPE-V EP *Almeida, Mendes & Kempster, 2016*

REFERENCES

1. Carding PN, Wilson JA, MacKenzie K, et al. Measuring voice outcomes: state of the science review. *J Laryngol Otol*. 2009;123:823–829.
2. Carding PN, Carlson E, Epstein R, et al. Formal perceptual evaluation of voice quality in the United Kingdom. *Logoped Phoniatr Vocol*. 2000;25:133–138.
3. Oates J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatr Logop*. 2009;61:49–56.
4. Wuyts FL, De Bodt MS, Van de Heyning PH. 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.
5. Speyer R. Effects of voice therapy: a systematic review. *J Voice*. 2008;22:565–580.
6. Hirano M. *Clinical Examination of Voice*. Vienna: Springer-Verlag; 1981.
7. American Speech-Language-Hearing Association. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), Special Interest Division 3, Voice and Voice Disorders. Available at: https://www.asha.org/uploadedFiles/members/divs/D3CAPEVprocedures.pdf . Accessed November 22, 2017.
8. Nemr K, Simões-Zenari M, Cordeiro GF, et al. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *J Voice*. 2012;26:812.e17–812.e22.
9. Aaronson N, Alonso J, Burman A, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002;11:193–205.
10. DeVon HA, Block ME, Moyle-Wright P, et al. A psychometric toolbox for testing validity and reliability. *J Nurs Scholarsh*. 2007;39:155–164.
11. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm*. 2008;65:2276–2284.
12. Franic DM, Bramlett RE, Bothe AC. Psychometric evaluation of disease specific quality of life instruments in voice disorders. *J Voice*. 2005;19:300–315.
13. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther*. 1996;18:979–992.
14. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull*. 1955;52:281–302.
15. Bassich CJ, Ludlow CL. The use of perceptual methods by new clinicians for assessing voice quality. *J Speech Hear Disord*. 1986;51:123–133.
16. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555–573.
17. Brinca L, Batista AP, Tavares AI, et al. The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *J Voice*. 2015;29:776.e7–776.e14.
18. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgements of dysphonic voice. *J Voice*. 2006;20:527–544.
19. Kreiman J, Gerratt BR. Validity of rating scale measurements of voice quality. *J Acoust Soc Am*. 1998;104:1598–1608.
20. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103–115.
21. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.
22. Kreiman J, Gerratt BR, Precoda K, et al. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35:512–520.
23. Maryn Y, Roy N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *J Soc Bras Fonoaudiol*. 2012;24:107–112.
24. Sofranko JL, Prosek RA. The effect of experience on classification of voice quality. *J Voice*. 2012;26:299–303.
25. Zraick RI, Wendel K, Smith-Olinde L. The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *J Voice*. 2005;19:574–581.
26. Jesus L, Barney A, Sá Couto P, et al. Voice Quality Evaluation Using CAPE-V and GRBAS in European Portuguese. In Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Florence, Italy, 2009b, 61−64.
27. Jesus L, Barney A, Santos R, et al. Universidade de Aveiro's voice evaluation protocol. In Proceedings of InterSpeech. Brighton, UK, 2009a, 971−974.
28. Karnell MP, Melton SD, Childes SJM, 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.
29. Kelchner LN, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24:441–449.
30. Mozzanica F, Ginocchio D, Borghi E, et al. Reliability and validity of the Italian version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Folia Phoniatr Logop*. 2013;65:257–265.
31. Nemr K, Simões-Zenari M, Souza GG, et al. Correlation of the Dysphonia Severity Index (DSI), Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), and gender in Brazilians with and without voice disorders. *J Voice*. 2015;25:765.e7–765.e11.
32. Núñez-Batalla F, Morato-Galán M, García-López I, et al. Validation of the Spanish adaptation of the consensus auditory-perceptual evaluation of voice. *Acta Otorrinolaringol Esp*. 2015;66:249–257.
33. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the consensus auditory-perceptual evaluation (CAPE-V). *Am J Speech Lang Pathol*. 2011;20:14–22.
34. De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:78–80.
35. Helou LB, Solomon NP, Henry LR, 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.
36. Kempster GB, Guerratt 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.
37. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
38. American Speech-Language-Hearing Association. Adult Hearing Screening. Available at: http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942721&section=Key_Issues. Accessed November 22, 2017.
39. Patel S, Shrivastav R. Perception of dysphonic vocal quality: some thoughts and research update—perspectives on voice and voice disorders. *ASHA*. 2007;17:3–7.
40. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
41. Harshbarger TR. *Introductory Statistics: A Decision Map*. 2nd ed. New York: Macmillan; 1977.