Reliability and Validity of the Turkish Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

*Esra Özcebe, *Fatma Esen Aydinli, *Tuğçe Karahan Tiğrak, *Önal İncebay, and †Taner Yilmaz, *†*Sıhhiye, Turkey*

Summary: Objectives. The main purpose of this study was to culturally adapt the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) to Turkish and to evaluate its internal consistency, validity, and reliability.

Materials and Methods. The Turkish version of CAPE-V was developed, and with the use of a prospective case-control design, the voice recordings of 130 participants were collected according to CAPE-V protocol. Auditory-perceptual evaluation was conducted according to CAPE-V and Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) scale by two ear, nose, and throat specialists and two speech and language therapists. The different types of voice disorders, classified as organic and functional disorders, were compared in terms of their CAPE-V scores.

Results. The overall severity parameter had the highest intrarater and inter-reliability values for all the participants. For all four raters, the differences in the six CAPE-V parameters between the study and the control groups were found to be statistically significant. Among the correlations for the comparable parameters of the CAPE-V and the GRBAS scales, the highest correlation was found between the overall severity-grade parameters. There was no difference found between the organic and functional voice disorders in terms of the CAPE-V scores.

Conclusions. The Turkish version of CAPE-V has been proven to be a reliable and valid instrument to use in the auditory-perceptual evaluation of voice. For the future application of this study, it would be important to investigate whether cepstral measures correlate with the auditory-perceptual judgments of dysphonia severity collected by a Turkish version of the CAPE-V.

Key Words: Auditory-perceptual assessment–CAPE-V–GRBAS–Dysphonia–Voice quality.

INTRODUCTION

Clinical voice evaluation starts with a case history interview and then proceeds to perceptual and instrumental assessments. Instrumental assessment of voice includes acoustic, aerodynamic, and other physiological measurements.[1] Perceptual assessment of voice involves auditory-perceptual judgment of voice quality, visual perceptual judgment of laryngoscopic examination, and the patients' judgment of their own voice problems.[2] These methods allow clinicians to describe the voice, understand the nature of the voice disorder, estimate the severity of dysphonia, and document either the changes over time or the treatment result.[1] Auditory-perceptual evaluation is a primary part of routine clinical voice assessment because of the ease and competency of the method.[3,4] In 1998, the British Voice Association stated that the three most commonly used formal rating scales were the Vocal Profile Analysis Scheme (VPA),[5] the Buffalo Voice Profile,[6] and the Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) scale.[7] The GRBAS scale, which is the most widely used of these three methods, was developed by the Japanese Society of Logopedics and Phoniatrics. In this method, G (grade) corresponds to the overall voice quality, R (roughness) corresponds to irregular fluctuations of frequency, B (breathiness) corresponds to the turbulence caused by air leakage, A (asthenia) corresponds to a hypokinetic and weak voice, and S (strain) corresponds to an effortful or hyperfunctional voice. In this method, the listener rates each voice quality feature by assigning a number ranging between 0 and 3 using a four-point Likert scale.[8] In a study in which the three auditory-perceptual methods were compared in terms of reliability, the GRBAS scale was proven to be a useful and fast clinical method that has good reliability for evaluating voice quality.[9] Although GRBAS is an easy, reliable, and valid auditory-perceptual assessment method, the narrow rating range is limited to 0–3 (normal, mild, moderate, and severe), making it difficult to rate subtle voice changes. In addition, because GRBAS does not have a specific protocol for data collection in terms of the variability of speech samples and the possible effects of task order, it is difficult to compare different raters' results across different studies.[10–12]

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) protocol was developed by the American Speech-Language-Hearing Association's Special Interest Division 3, Voice and Voice Disorders, and was adopted by consensus at a conference in 2002 at the University of Pittsburgh.[13] The CAPE-V protocol has a perspective different from other voice assessment methods; the protocol was developed by addressing the psychoacoustic and psychophysical issues related to human perception and scaling. The protocol includes gathering voice samples of patients representing (1) sustaining vowels /a/ and /i/, (2) reading sentences, and (3) conversational speech. The sentences are specifically developed based on different phonetic contexts. The CAPE-V protocol also provides for a standard data

Accepted for publication November 21, 2017.

The study, which includes the preliminary results of the present study with the title "Reliability and Validity of the Turkish Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)/Preliminary Results," had been accepted for presentation as a poster at the Meeting of the Voice Foundation—45th Annual Symposium, 2016.

From the *Department of Speech and Language Therapy, Hacettepe University Faculty of Health Sciences, Sıhhiye, Ankara, Turkey; and the †Department of Otolaryngology-Head and Neck Surgery, Division of Laryngology and Phonosurgery, Hacettepe University Faculty of Medicine, Sıhhiye, Ankara, Turkey.

Address correspondence and reprint requests to Esra Özcebe, Department of Speech and Language Therapy, Hacettepe University Faculty of Health Sciences, Sıhhiye, 06100 Ankara, Turkey. E-mail: eozcebe@gmail.com

Journal of Voice, Vol. 33, No. 3, pp. 382.e1–382.e10
0892-1997

https://doi.org/10.1016/j.jvoice.2017.11.013

collection system. After listening to the samples, voice quality is rated by scoring six characteristics: overall severity, roughness, breathiness, strain, pitch, and loudness. All of these parameters are labeled on a 100-mm visual analog scale (a horizontally oriented line), which has open-ended anchors. The rater marks the point that shows voice quality deviance.[13] Pitch and loudness changes, which are not rated in the GRBAS protocol, are also evaluated separately.

The CAPE-V protocol has been used in many studies that have investigated the perceptual evaluation of voice.[11–15] The concurrent validity of the CAPE-V was established by Zraick et al.[16] In that study, the intra- and inter-reliability coefficients were found to be slightly higher for the CAPE-V protocol than for the GRBAS scale. In addition, CAPE-V was found to be more useful for detecting small changes in voice.[11]

The CAPE-V protocol has been proven to be a reliable instrument in studies investigating the use of original developed language. One of the most important advantages of the CAPE-V protocol is that it includes standard continuous speech samples. Continuous speech assessment of the perceptual assessment of voice provides clinicians with a more realistic evaluation of a voice disorder. However, the assessment may be affected by linguistic, phonetic, and cultural factors.[17] Many studies have investigated the cross-linguistics aspects of voice quality.[18,19] When taking the Turkish language into account, we focused on two studies. In Uygun et al's study,[20] which examined the frequency of hard glottal attacks (HGAs) among healthy and dysphonic Turkish participants, the researchers found that there were fewer HGAs in both groups in comparison with English language speakers. Uygun et al[20] hypothesized that structural differences between the Turkish and English languages that are related to "stress" may be responsible for this finding. In Turkish, the stress is always on the last syllable, whereas in English, the stress may fall on the first, last, or second-to-last syllable. In Bahmanbiglu et al's study,[19] voice quality was evaluated in bilingual participants who speak Farsi and Turkish. On a long-term average spectrum, Bahmanbiglu et al calculated the mean spectral energy and the spectral tilt. Bahmanbiglu et al[19] found that in Turkish, sentences were produced with a higher laryngeal tension and a breathier voice quality. The researchers hypothesized that this difference may be due to the contrasting resonance patterns of the languages, which should be given greater consideration in the literature.

To use the CAPE-V protocol in clinical voice evaluations, CAPE-V versions of Italian,[12] Portuguese,[21] and Spanish[22] have been developed. In addition to developing language-specific sentences in each language, the reliability and validity of CAPE-V should also be investigated. It is still controversial as to whether voice quality is a universal property of voice or is language-dependent.[11,23–25] In the study conducted by Yamaguchi et al,[24] Japanese and American listeners were the evaluators and the GRBAS method was used. Yamaguchi et al indicated that linguistic factors did not affect the audioperceptual evaluation of the grade parameter, but found that the asthenia and strain parameters were different, depending on the speaker's language. In a study with a similar methodology[16,26] the breathy and roughness parameters were rated differently for Cantonese and English speakers. The results of these studies support the idea that audioperceptual assessment is affected by linguistic and cultural factors; based on the specific properties of different languages, the same protocol may have different results. Culturally adapting CAPE-V to Turkish not only provides clinicians with the ability to conduct cross-cultural studies but also provides information about the influence of language on the continuous speech assessment of the audioperceptual evaluation of voice.[16]

The main purpose of the present study was to culturally adapt the CAPE-V to Turkish and to evaluate its internal consistency, validity, and reliability. The second purpose of the present study was to compare the CAPE-V scores of functional and organic voice disorders.

MATERIAL AND METHODS

The study participants were recruited from patients who visited Hacettepe University's ear, nose, and throat department and speech-language pathology department from September 2015 to November 2016. All the evaluations were carried out in the ear, nose, and throat department and the speech language pathology department at Hacettepe University Hospital. The study used a prospective case-controlled design methodology. The study obtained approval from Hacettepe University's Non-invasive Clinical Research Ethics Committee on April 11, 2015 (Project No: G0 15/675).

Participants

Inclusion criteria

Voice recordings from 140 participants were reviewed and the records that did not comply with the recording protocol were excluded from the study. Consequently, the voice recordings of 130 participants were included in the study. The participants were placed into two groups: a study group and a control group. The study group (n = 76) consisted of patients diagnosed with vocal nodules (n = 22), muscle tension dysphonia (MTD) (n = 17), sulcus vocalis (n = 13), mutational falsetto (n = 6), polyp (n = 5), cyst (n = 4), Reinke edema (n = 3), unilateral vocal fold paralysis (n = 3), localized Reinke edema (n = 2), and spasmodic dysphonia (n = 1) (Table 1). The participants in the study group should have a certain diagnosis of voice disorder and should have

TABLE 1.
Distribution of the Diagnoses in the Study Group

Diagnosis	n	%
Nodule	22	28.94
MTD	17	22.36
Sulcus	13	17.10
Mutational falsetto	6	7.89
Polyp	5	6.57
Cyst	4	5.26
Paralysis	3	3.94
Reinke edema	3	3.94
Localized Reinke edema	2	2.63
Spasmodic dysphonia	1	1.31
Total	76	100.0

TABLE 2.
Demographic Characteristics (Gender and Age Distributions) of all Participants in the Control Group and in the Study Group

	All Participants			Control Group			Study Group		
	n	%	Mean Age	n	%	Mean Age	n	%	Mean Age
Female	87	66.9	31.6	37	68.5	30.2	49	64.5	32.8
Male	43	33.1	35.2	17	31.5	34.0	27	35.5	34.8
Total	130	100.0	33.4	54	100.0	32.1	76	100.0	33.8

good reading skills (a minimum primary school education). They should not have been previously treated for dysphonia; should not have any neurologic disease, upper aerodigestive tract malignancy, or hearing loss; and should not be using any medication. The control group (n = 54) consisted of individuals with no voice disorder. The participants in the control group met the following conditions: no voice complaints either before or on the evaluation day, no neurologic or systemic disease that could affect the voice, no cold on the evaluation day, had not smoked cigarettes for a minimum of 5 years, had a 7 or lower score on the Voice Handicap Index-10 form, and had good reading skills (a minimum primary school education). In addition, two SLTs with experience in voice disorders confirmed audioperceptually that the participants' voices were normal.

Demographic characteristics of the participants

The age range was 18–69 years for both groups; the mean age was 33.84 years for the study group and 32.11 years for the control group, and there was no statistically significant difference between the groups ($P = 0.523$). In the study group, n = 49 (65.3%) of the participants were female; in the control group n = 37 (69.8%) were female. The gender ratios were similar between the groups ($P = 0.328$). Table 2 shows the demographic characteristics of all participants, the control group, and the study group.

Voice evaluations

Ear, nose, and throat evaluation

Laryngoscopic examination was performed using flexible 3.7-mm diameter steerable fiber-optic laryngoscopy (Optim, Sturbridge, MA) and rigid video laryngostroboscopy using a Kay Pentax digital strobe (Kay Pentax, Lincoln Park, NJ) and Kay Pentax Rls 9100 B equipment (Key Elemetrics, Lincoln Park, NJ). Diagnosis for each patient was determined by a voice council team consisting of ear, nose, and throat specialists (ENTs) and speech and language therapists (SLTs). A laryngologist examined the laryngeal images to confirm the diagnosis, and ENTs and SLTs discussed the patients' objective and subjective voice evaluation results.

Audioperceptual data gathering

All the participants' voice recordings were obtained in accordance with the CAPE-V protocol (sustaining /a/ and /i/, sentences, and conversational speech).[13] Conversational speech was gathered for a maximum of 2 minutes, and the patients were informed that their personal information would not be used in the study. The evaluation was done in a room in which the environmental noise was <50 dB.[27] Voice recordings were captured using the *Analysis of Dysphonia of Speech and Voice* program (CSL Model 4500 equipment, Kay Elemetrics Group, Lincoln Park, NJ, USA).[28] Voice recordings were obtained using the default sampling rate 25.000 and were saved in .wav format. During the recordings, a Micromic C520 headset microphone was used, maintaining a distance of 5 cm and a 45° angle to the mouth. All the recordings were stored in a universal serial bus device.

Developing the Turkish version of CAPE-V and an audioperceptual evaluation of voice

Permission was obtained from the American Speech-Language-Hearing Association copyright department to develop a Turkish version of the CAPE-V protocol. The team, including one linguist and one SLT, developed the Turkish CAPE-V sentences by following the phonetic rules described in the CAPE-V application protocol.[13]

Audioperceptual evaluation

Audioperceptual evaluation was done by two ENTs and two SLTs according to the GRBAS and CAPE-V protocols.[13] All the raters had a minimum of 5 years of experience in the voice disorder field. All of the raters completed their evaluations in two separate sessions with a minimum of 48–72 hours between evaluations. Two of the raters (rater 1 and rater 2) first rated the voices based on GRBAS, and two of the raters (rater 3 and rater 4) first rated voices based on CAPE-V.

The raters listened to the recordings in a free-field environment to minimize background noise and to provide a calmer setting. All raters used a MacBook Air (Apple Inc, Cupertino, CA) while listening to the samples. The raters were allowed to adjust the volume to a comfortable, consistent level and were allowed to play back the recordings as often as they wanted. The sound file names were coded by numbers, which did not include any name or group information.

Anchor samples of healthy and dysphonic voices were used for familiarization with the protocol.[16] Before each rating session, the raters were asked to listen to the four anchor samples. The four familiarization sample voices were mastered and labeled as recordings 1–4. One male and one female normal voice, followed by the two dysphonic voices (one considered to be mild dysphonia and another judged to be severely dysphonic) were included. These voices were judged according to a four-point Likert scale (1 = normal, 2 = mild dysphonia, 3 = moderate dysphonia, and 4 = severe dysphonia) by an SLT following the procedures used in Zraick et al's study.[16] All four voice samples used for the task familiarization did not belong to the participants in the present study.

After listening to the anchor samples, the study and control groups' recordings were randomly given to them. All of the raters were blind to the all participants and to the participants' diagnoses in the study group. The raters did not receive any information about the participants' names. For the intrarater reliability analysis, 15.50% (n = 20) of the recordings were obtained after a minimum of 1 week; the samples that were used to determine inter-rater reliability were chosen randomly and included samples from both the control and the study groups.

The raters marked the CAPE-V evaluation form, calculating their scores for the six voice quality parameters; each rater marked thick lines (using paper and pencil) along the 100-mm horizontal line for each designated parameter.

Validity analysis

To determine the validity of the Turkish CAPE-V, the scores of the study and the control groups were compared. In addition, a discriminant function analysis was conducted. To determine concurrent validity, the results of the GRBAS evaluation were compared with the results of the CAPE-V evaluation.

Statistical analysis

Statistical tests were performed using *Statistical Package for the Social Sciences* (SPSS), Version 18 software (SPSS, Inc, Chicago, IL). Pearson chi-square test was used to determine differences between the study and control groups in terms of gender, and the Mann-Whitney *U* test was used to determine differences in terms of age. The intra- and inter-rater reliability of the CAPE-V was evaluated using a two-way mixed-effect model, and intraclass correlation coefficients (ICCs) were determined. Intrarater reliability was determined by each of the raters listening to 20 voice samples (15.50%). The Mann-Whitney *U* test was used to compare the CAPE-V values of the study group with those of the control group, and to compare different group of voice disorders. To control for the increased risk of type I errors resulting from the large number of comparisons assessed by the Mann-Whitney tests, Bonferroni corrections were performed and a per-comparison alpha level was set at $P = 0.001$. The degree of association between the CAPE-V and GRBAS comparable parameters was determined using Spearman correlations.[29] Discriminant function analysis and classification were conducted to determine whether overall severity could predict study groups (study and normal). The diagnoses were classified as functional voice disorders (mutational falsetto and MTD and as structural voice disorders (nodules, sulcus, polyps, Reinke edema, vocal fold paralysis, and spasmodic dysphonia),[30] and CAPE-V scores of functional and structural voice disorders were compared by the independent samples Student *t* test.

RESULTS

The voice recordings of 54 participants in the control group and 76 participants in the study group were rated according to the CAPE-V and GRBAS protocols. The six basic parameters of CAPE-V (overall severity, roughness, breathiness, strain, pitch, and loudness) were included in the evaluation.

Reliability analysis

Intrarater reliability

Intrarater reliability was determined by the ICCs for each of the CAPE-V parameters and was calculated for each rater separately. As seen in Table 3, the ICC values for all the raters were higher than .92 for the overall severity parameter. For the roughness parameter, only one rater's ICC value was lower than .90, and the ICC values for all of the other raters were higher than .95. Similarly, only one rater (rater 4) had an ICC value lower than .90; the ICC values for all the other raters were equal to or higher than .90. Thus, it can be clearly stated that the strain parameter had the lowest ICC values; all the ICC values for this parameter were lower than .90 (in the range of .76–.86). All of the ICC values for the pitch and loudness parameters were higher than .85, which indicates high correlation.

Inter-rater reliability

The inter-rater reliability results for all raters are shown in Tables 4 and 5. In the overall group reliability analysis (Table 4), the highest ICC value was .90 for the overall severity parameter. The ICC values were higher than .80 for four parameters: roughness, breathiness, loudness, and pitch. The lowest ICC value was .80 for the strain parameter. When inter-rater reliability was investigated separately for the control and the study groups (Table 5),

TABLE 3.
Intrarater Reliability Analysis of the CAPE-V Using ICC Analysis for all Raters

			95% Confidence Interval	
Parameter	Rater	Intraclass Correlation Coefficient	Lower Limit	Upper Limit
Overall severity	Rater 1	.96	.92	.98
	Rater 2	.98	.97	.99
	Rater 3	.93	.83	.97
	Rater 4	.96	.91	.98
Roughness	Rater 1	.96	.90	.98
	Rater 2	.98	.95	.99
	Rater 3	.89	.74	.96
	Rater 4	.95	.88	.98
Breathiness	Rater 1	.93	.82	.97
	Rater 2	.98	.95	.99
	Rater 3	.90	.77	.96
	Rater 4	.89	.74	.95
Strain	Rater 1	.76	.39	.90
	Rater 2	.86	.64	.94
	Rater 3	.83	.58	.93
	Rater 4	.86	.66	.94
Pitch	Rater 1	.89	.74	.96
	Rater 2	.95	.89	.98
	Rater 3	.88	.70	.95
	Rater 4	.99	.97	.99
Loudness	Rater 1	.86	.66	.94
	Rater 2	.97	.93	.99
	Rater 3	.93	.84	.97
	Rater 4	.97	.94	.99

TABLE 4.
Inter-rater Reliability Analysis of the CAPE-V Using ICC Analysis in the Overall Group

Parameter	Intraclass Correlation Coefficient	95% Confidence Interval	
		Lower Limit	Upper Limit
Overall severity	.90	.87	.92
Roughness	.81	.75	.85
Breathiness	.84	.80	.88
Strain	.80	.74	.85
Loudness	.81	.75	.86
Pitch	.88	.84	.91

the highest ICC values for the overall severity parameter were .88 and .91 for the control and the study groups, respectively. In the study group, all the other ICC values were higher than .80; the roughness parameter had the lowest value (.84). In the control group, all the ICC values were somewhat lower than the values in the study group. In the control group, all the ICC values were higher than .70; the breathiness parameter had the lowest ICC value (.72).

Validity analysis

For all four raters, the median and quartile values of the CAPE-V and the differences in the six CAPE-V parameters between the study and the control groups are shown in Table 6. The results show that the raters had a tendency of scoring 0 for the healthy voices. The overall severity and roughness parameters had the highest third quartile score of 17.5 in the control group. In the study group, it was observed that the median scores were frequently around the moderate deviance area, and the strain parameter had the highest third quartile value with a score of 70. The Mann-Whitney *U* test results showed a significant difference for the all parameters between the two groups ($P < 0.01$).

A discriminant function analysis was conducted to determine whether the overall severity level could predict in the study groups (control and study). The overall Wilks lambda was significant, $\Lambda = .61$, chi-square $(1, N = 130) = 58.37$, $P < 0.05$, which indicates that the overall severity differed between the control and the study groups. Classification results showed that 85.1% of the participants in the present study were correctly classified. Additionally, the sensitivity was 83%, the specificity was 89%, the positive predictive value was 93%, and, finally, the negative predictive value was 75%. The discriminant function analysis and classification results showed that the overall severity subscale of CAPE-V is an adequate tool to differentiate between the control and the study groups.

The diagnoses were classified as functional voice disorders (mutational falsetto and MTD) and as structural voice disorders (nodules, sulcus, polyps, Reinke edema, vocal fold paralysis, and spasmodic dysphonia).[30] Then CAPE-V scores of each parameter were compared between the functional and structural voice disorders. Before comparison, normality assumptions were tested and it was observed that scores were normally distributed within these groups. Additionally, Levene test results indicated that variances in both groups were homogenous ($F = 1.31$, $P > 0.05$). Hence, an independent samples *t* test was used to compare groups. Statistical analysis did not show a significant difference between the two groups for all parameters of CAPE-V ($P > 0.05$). Sample sizes, means and standard deviations, and *P* values of the analyses are presented in Tables 7 and 8.

In Table 8, it is seen that all the *P* values are higher than .05, which indicated that there is no difference in any of the CAPE-V parameters between the organic and structural voice disorders.

Concurrent validity

According to the correlations for the comparable parameters of the CAPE-V and GRBAS scales (Table 9), grade-overall severity parameters were found to be highly correlated (.80) and the other parameters were found to be moderately correlated.

DISCUSSION

Auditory-perceptual voice evaluation has many advantages over physiological measurements and is easy to implement. However, this method is only valuable if the evaluation method used is valid and reliable.[15,31] Many factors may affect the results of an auditory-perceptual evaluation. These include the internal factors related to the listener, such as the listener's background and experience, the consensus training provided for raters,[32,33] and the type of rating scale used.[31] The CAPE-V protocol is widely used in auditory-perceptual rating; however, linguistic and cultural factors may substantially affect the evaluation process. For

TABLE 5.
Inter-rater Reliability Analysis of the CAPE-V Using ICC Analysis in Both the Control Group and the Study Group

Parameter	Intraclass Correlation Coefficient	95% Confidence Interval		Intraclass Correlation Coefficient	95% Confidence Interval	
		Lower Limit	Upper Limit		Lower Limit	Upper Limit
Overall severity	.88	.82	.93	.91	.88	.94
Roughness	.83	.73	.90	.84	.84	.89
Breathiness	.72	.56	.67	.89	.84	.92
Strain	.84	.74	.90	.87	.81	.91
Loudness	.77	.62	.86	.88	.83	.92
Pitch	.83	.73	.90	.89	.84	.92

TABLE 6.
Median and Quartile Values of CAPE-V and Comparison of the Control Group and the Study Group for all the Six Parameters of CAPE-V

CAPE-V	Control Group		Study Group		
Parameter	Median	Percentiles (Q1–Q3)	Median	Percentiles (Q1–Q3)	*P*
OS					
R1	0	0–10	40	20–50	0.000*
R2	5	0–17.5	45	30–70	0.000*
R3	0	0–0	25	12.5–50.0	0.000*
R4	5	0–10	39	20.0–58.7	0.000*
R					
R1	0	0–7.5	25	10–50	0.000*
R2	0	0–17.5	40	20–60	0.000*
R3	0	0–0	25	.7–50.0	0.000*
R4	5	2–10	34.5	18–55	0.000*
B					
R1	0	0–0	10	0–25	0.000*
R2	0	0–5	30	10–50	0.000*
R3	0	0–0	25	0–28.5	0.000*
R4	0	0–2	15	5–30	0.000*
St					
R1	0	0–0	20	10.0–37.5	0.000*
R2	0	0–13.7	50	30–70	0.000*
R3	0	0–0	25	10–50	0.000*
R4	0	0–0	25	8.2–50.0	0.000*
P					
R1	0	0–0	20	6.2–40.0	0.000*
R2	0	0–10	50	20–70	0.000*
R3	0	0–0	25	4.7–50	0.000*
R4	0	0–0	25	15–55	0.000*
L					
R1	0	0–0	15	0–30	0.000*
R2	0	0–5	40	20–60	0.000*
R3	0	0–0	25	0–50	0.000*
R4	0	0–0	24.5	6.2–40.0	0.000*

* $P < 0.01$.

Abbreviations: B, breathiness; L, loudness; OS, overall severity; P, pitch; Q1, first quartile (25th percentile); Q3, third quartile (75th percentile); R, roughness; St, strain.

TABLE 7.
Comparison of CAPE-V Scores Between Structural Voice Disorders and Functional Voice Disorders: The Mean Scores and Standard Deviations

		Overall Severity		Roughness		Breathiness		Strain		Pitch		Loudness	
	N	Mean (Score)	SD	Mean (Score)	SD	Mean (Score)	SD	Mean (Score)	SD	Mean (Score)	SD	Mean (Score)	SD
Structural voice disorders	53	37.45	28.09	36.00	25.52	34.87	31.00	40.36	29.52	35.16	27.04	40.38	37.03
Functional voice disorders	23	36.43	33.68	34.07	29.74	37.12	27.91	38.71	27.06	37.88	31.25	38.04	36.54
Total	76												

Abbreviation: SD, standard deviation.

TABLE 8.
Comparison of CAPE-V Scores Between Structural Voice Disorders and Functional Voice Disorders: The *P* Values of the Independent Samples *t* Test for Each Parameter

Overall Severity	Roughness	Breathiness	Strain	Pitch	Loudness
.072	.06	.081	.062	.07	.084

instance, it is shown that HGAs are less frequently produced in Turkish compared with English.[20] A CAPE-V protocol that considers linguistic and cultural factors should be developed for each language, and its internal consistency, validity, and reliability should be determined. Zraick et al[16] demonstrated the concurrent validity of the CAPE-V protocol that was originally developed for English. Italian,[12] Portuguese,[21] and Spanish[22] versions of CAPE-V have also been developed, and their reliability and validity have been investigated.

A review of the literature shows that, when inter-rater reliability is considered, the overall severity parameter has the highest correlation values across studies.[11,12,14,16] Except for the correlation value of .76 obtained by Zraick et al,[16] the values for the overall severity parameter in other studies were all higher than .90.[11,12,14] Zraick et al[16] used 21 raters and involved fewer participants (n = 59), so these factors could have influenced the lower correlation values that were obtained. In the present study, in agreement with the findings reported in the literature, the overall severity parameter was found to have the highest correlation value (.90) in the overall group; this value was .88 in the control group and .91 in the study group.

Because some studies included only dysphonic voices, the correlation values in these studies were based on whether they had a control group. Table 5 presents the values of the groups separately. When the parameters were analyzed in terms of the lowest correlation value for inter-rater reliability, rating the strain parameter was found difficult.[11,14,34] In the present study, consistent with the literature, the lowest value (.80) was found for the strain parameter in the overall group. In addition, all the inter-rater values were found to be somewhat lower in the control group for each parameter. As some researchers have hypothesized, this result may be associated with continuous speech characterized by intermittent vocal fry or roughness[12,16]; however, in the present study, it is pointed that this result can be explained by the reduced range of variation present in the control group, which is rated along a smaller proportion of the scale relative to the study group. The strain parameter had the lowest correlation value probably because of the population used in the present study. We included participants with 10 different vocal diagnoses: vocal nodules (n = 22), MTD (n = 17), sulcus vocalis (n = 13), mutational falsetto (n = 6), polyp (n = 5), cyst (n = 4), Reinke edema (n = 3), unilateral vocal fold paralysis (n = 3), localized Reinke edema (n = 2), and spasmodic dysphonia (n = 1). However, there was no equal number of participants for each diagnosis, so this may have affected the reliability of the results. For example, spasmodic dysphonia, which is frequently associated with vocal effort,[35] may have caused the lower reliability values for the strain parameter. In addition, the present study's participants were patients who were directed for voice therapy primarily, so the assumption was that there were fewer patients with severe dysphonia. In the literature, rating the specific voice parameters for mild-to-moderate dysphonia has been shown to be more challenging.[36]

In the present study, the intrarater reliability values were high; they were in the range of .86–.98 for all the CAPE-V parameters, except for strain. Similarly, in the literature, intrarater correlation values in the range of .92–.98 were found[11,12,14] for the overall severity parameter. In addition, the roughness[12] and breathiness[16] parameters had high values. In the present study, the correlation value of the strain parameter indicated a high correlation but had the lowest intrarater values in the range of .76–.86. Zraick et al[16] and Kelchner et al[34] reported the lowest correlation values for the strain parameter, which is similar to our result. Mozzanica et al[12] reported the lowest correlation value for the loudness parameter (Table 9).

In summary, consistent with the results reported in previous studies, overall severity, roughness, and breathiness had higher intrarater reliability correlation values, whereas strain, pitch, and loudness somewhat had low intrarater reliability correlation values. Similarly, in many studies investigating the intra- and inter-rater reliability of the GRBAS scale, overall severity, roughness, and breathiness were found to have well to high reliability, whereas the strain parameter had the lowest intrarater reliability values.[9,11,25] These findings indicate that the auditory-perceptual evaluation of voice quality by CAPE-V is less likely to be affected by linguistic properties; however, the methodology used in the present study is assumed to be insufficient to verify this claim. The hypothesis is that studies that were structured by considering the similarities and discrepancies between language couples may help reveal the role of language in the auditory-perceptual assessment of voice.

The results of the present study show that the Turkish version of CAPE-V has high intrarater and inter-rater reliability. In addition, results of the discriminant function analysis and classification showed that CAPE-V is an adequate tool to differentiate between study and control groups. Although some studies hypothesized that a listener's background affects his or her auditory-perceptual analysis,[37,38] our study did not support this hypothesis. In the current study, two of the raters were ENTs and two were SLTs, and all of them had a minimum of 5 years' experience in dealing with voice disorders. This finding can be the result of the anchor samples that the raters listened to before the start of the rating process. All the raters listened to four voice samples (including normal and pathologic voices) before rating the voice recordings used in the study. However, as a limitation of the present study, the education provided in this study was a self-familiarization protocol that includes only definitions and written instructions on the CAPE-V rating procedure.

TABLE 9.
Ranges of Intra- and Inter-reliability for the CAPE-V and its Correlation Values With the GRBAS Across Studies

Data Reported	Mozzanica et al (2014)			Nemr et al (2012)	Zraick et al (2011)	Karnell et al (2007)	Present Study		
Raters and experience	SLPs (n = 3), >5 y			SLPs (n = 3), >5 y	SLPs (n = 21), >5 y	SLPs (n = 4), >5 y	2 SLPs and 2 ENTs, >5 y		
Language	Italian			Portuguese	English	English	Turkish		
Sample (n)	n = 200 (control = 120, study = 80)			n = 60	n = 59	n = 34	n = 130 (control = 54, study = 76)		
Intrarater reliability									
Statistics (group)	ICCs			ICCs	ICCs	Spearman correlations (study group)	ICCs		
Minimum	.80 (L)				.35 (S)	.88	.76 (St)		
Maximum	.92 (OVS, R)			.923–.985 (G)	.82 (B)	.93 (OVS)	.98 (OVS, R, and B)		
Inter-rater reliability									
Statistics	ICCs (overall)	Control group	Study group	ICCs (study group)	ICC (study group)	Spearman correlations (study group)	Overall	Control G.	Study group
Minimum	.76 (St)	.78 (L)	.82 (L)	.828 (St)	.28 (pitch)	.86	.80 (St)	.72(B)	.84 (R)
Maximum	.92 (OVS)	.93 (OVS)	.91 (OVS, R)	.911 (OVS)	.76 (OVS)	.93 (OVS)	.90 (OVS)	.88(OVS)	.91 (OVS)
Correlation statistics parameter (value)	Spearman correlations			Spearman correlations	Multiserial correlations	Spearman correlations	Spearman correlations		
Minimum	.79 (St-St)			.84(OVS-G)	76 (R-R)	.89 (B-B)	.62 (R-R)		
Maximum	.92 (OVS-G)				.80 (OVS-G)	.95 (OVS-G)	.80 (OVS-G)		

Abbreviations: B, breathiness; G, grade; L, loudness; R, roughness; S, severity; SLP, speech-language pathologist; St, strain; ENT, ear, nose, and throat specialist; OVS, overall severity.

TABLE 10.
Average Spearman Correlations Between Comparable CAPE-V and GRBAS Scales

CAPE-V	GRBAS	Correlation Mean *rs*
Overall severity	Grade	.80
roughness	Roughness	.62
breathiness	Breathiness	.67
strain	Strain	.68

Despite the healthy and dysphonic voice samples included, there was no single training protocol agreed on. The assumption is that the inter-rater reliability values would have been higher if a consensus training was provided before the auditory-perceptual evaluation.[32,39]

When the correlation between the GRBAS and CAPE-V scales was considered, the overall severity-grade of the parameters was found to have a high correlation value, and all the other parameters had moderate correlation values (Table 10). Other studies in the literature have shown overall severity-grade correlations ranging from .80 to .95.[11,12,14,16,22] Nemr et al,[14] Zraick et al,[16] and Núñez-Batalla et al[22] reported values close to .80, and Karnell et al[11] and Mozzanica et al[12] reported values higher than .90. In the present study, the values were lower than those reported by Mozzanica et al[12] and Karnell et al.[11] One possible explanation for this could be the difference in the timing of the application of the two scales. Karnell et al[11] applied both scales at the same time, but in the present study, there was a 48- to 72-hour time difference between the applications of each of the two scales. A possible explanation for the difference between the results reported by Mozzanica et al[12] and those of the present study may be the variations in the listeners' and the raters' clinical and educational backgrounds, the consensus training of the raters, and the different diagnoses and dysphonia severities in the study groups. When the lowest values were compared in terms of the correlation between the two scales, roughness (the present study and that of Zraick et al[16]), breathiness,[12,22] and strain[12] were the parameters that had the lowest correlation values. Núñez-Batalla et al[22] reported a correlation of .612 between the two breathiness scales. This result could be attributed to the differences between the studies in terms of the number of participants involved, the variability of the patients' diagnoses and dysphonia severities, and the timing of the application of the two scales.[32,33,36,37]

In the present study, comparing the CAPE-V scores between the functional and the organic types of voice disorders did not reveal any difference for any of the CAPE-V parameters. These findings support the idea that different kinds of voice disorders can result in similar perceptual properties.[40] In the study of Mozzanica et al,[12] research classified patients (n = 80) into groups according to their vocal pathologies and compared the CAPE-V scores for each parameter. Mozzanica et al included participants with the diagnosis of polyp, Reinke edema, nodule, MTD, unilateral vocal fold paralysis, and scar. They reported significant lower scores for the participants with polyps on the loudness parameter. Evaluation of loudness and pitch parameters is one of the most different aspects of the CAPE-V protocol compared with the GRBAS protocol. Evaluating pitch as a seperate parameter may be very useful of following patients with voice disorders pitch is mainly considered, for example, mutational falsetto or transgender voice. However, the results of the present study did not support the suggestion that the CAPE-V protocol may have specific auditory-perceptual profiles for different types of voice disorders.

CONCLUSIONS

In the present study, the Turkish version of CAPE-V has been proven to be a reliable and valid instrument. The results of the intra- and inter-rater reliability and validity analyses were in agreement with the findings of similar studies in the literature. In general, the overall severity parameter had the highest intra- and inter-reliability values and had a high correlation with the grade parameter, whereas the strain parameter had the lowest reliability values.

Culturally adapting CAPE-V to Turkish not only provides clinicians with the ability to conduct cross-cultural studies audioperceptually but also enables them to conduct studies instrumentally. It is thought that, for the future application of the present study, it would be important to investigate whether cepstral measures correlate with the auditory-perceptual judgments of dysphonia severity collected by a Turkish version of the CAPE-V, and a Turkish version of the Cepstral Spectral Index of Dysphonia should also be developed.[41,42]

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at doi:10.1016/j.jvoice.2017.11.013.

REFERENCES

1. Colton RH, Casper JK, Leonard R. *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*. 3rd ed. Baltimore: Lippincott Williams & Wilkins; 2011.
2. Carding P. The speech and language therapist's assessment of the dysphonic patient. In: Freeman M, Fawcus M, eds. *Voice Disorders and Their Management*. 3rd ed. London: Wiley; 2001:69–80.
3. Hirano M. Objective evaluation of the human voice: clinical aspects. *Folia Phoniatr Logop*. 1989;144:41–89.
4. Barstiers B, De Bodt M. Assessment of voice quality: current state-of-the-art. *Auris Nasus Larynx*. 2015;42:183–188.
5. Laver J, Wirz S, Mackenzie J, et al. A perceptual protocol for the analysis of vocal profiles. In: *Work in Progress*, Vol. 14. University of Edinburgh, Department of Linguistics; 1981:139–155.
6. Carding P, Carlson E, Epstein R, et al. Formal perceptual evaluation of voice quality in the United Kingdom. *Logoped Phoniatr Vocol*. 2000;25:133–138.
7. Isshiki N, Okamura H, Tanabe M, et al. Differential diagnosis of hoarseness. *Folia Phoniatr Logop*. 1969;21:9–19.
8. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *J Voice*. 2004;18:299–304.
9. Webb A, Carding P, Deary I, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol*. 2004;261:429–434.

10. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am*. 2007;122:2354–2364.
11. Karnell MP, Melton SD, Childes JM, 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.
12. 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.
13. Kempster GB, Gerratt BR, Abbott KV, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
14. Nemr K, Simões-Zenari M, de Souza GS, 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;30:765.e7–765.e11.
15. Solomon NP, Helou LB, Stojadinovic A. Clinical versus laboratory ratings of voice using the CAPE-V. *J Voice*. 2011;25:e7–e14.
16. Zraick RI, Kempster GB, Connor NP. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20:14–22.
17. Ghio A, Cantarella G, Weisz F. Is the perception of dysphonia severity language-dependent? A comparison of French and Italian voice assessments. *Logoped Phoniatr Vocol*. 2015;40:36–43.
18. Ordin M, Mennen I. Cross-linguistic differences in bilinguals' fundamental frequency ranges. *J Speech Lang Hear Res*. 2017;60:1493–1506.
19. Bahmanbiglu SA, Mojiri F, Abnavi F. The impact of language on voice: an LTAS study. *J Voice*. 2017;31:249, e9-e12.
20. Uygun MN, Aydınlı FE, Aksoy S, et al. Turkish standardized reading passage for the evaluation of hard glottal attack occurrence frequency. *J Voice*. 2017;doi:10.1016/j.jvoice.2017.03.004.
21. de Almeida SC. A thesis submitted in partial fulfillment of the requirement for the degree of Master in Science at the Health Science School of Polytechnic Institute of Setúba. Validity and reliability of the 2nd European Portuguese version of the "Consensus Auditory-Perceptual Evaluation of Voice" (II EP CAPE-V). Available at: https://comum.rcaap.pt/bitstream/10400.26/17609/1/II%20CAPE%20EP%20v%20final_submetida%2020_12_2016.pdf. Accessed April 2, 2017.
22. Núñez-Batalla F, Morato-Galán M, García-López I, et al. Validation of the Spanish adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Acta Otorrinolaringol Engl*. 2015;66:249–257.
23. Kent RD, Ball MJ. *Voice Quality Measurement*. 1st ed. Singular; 2000.
24. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatr Logop*. 2003;55:147–157.
25. 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-e22.
26. Yiu EML, Murdoch B, Hird K, et al. Cultural and language differences in voice quality perception: a preliminary investigation using synthesized signals. *Folia Phoniatr Logop*. 2008;60:107–119.
27. Titze IR. Summary statement. Workshop on Acoustic Voice Analysis. National Center for Voice and Speech, Wendell Johnson Speech and Hearing Center, The University of Iowa, Iowa City, Iowa; February 17th, 18th, 1994:26–32.
28. Pentax Medical. ADSV manual, medical. 2017. Available at: http://www.pentaxmedical.com/pentax/en/92/1/Downloads-for-PDF. Accessed May 5, 2017.
29. Field A. *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications Ltd.; 2009.
30. Baker J. Functional voice disorders: clinical presentations and differential diagnosis. *Handb Clin Neurol*. 2017;139:389–405.
31. Oates J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatr Logop*. 2009;61:49–56.
32. Iwarsson J, Petersen NR. Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *J Voice*. 2012;26:304–312.
33. 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.
34. Kelchner LN, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24:441–449.
35. Houtz RD, Roy N, Merrill RM, et al. Differential diagnosis of muscle tension dysphonia and adductor spasmodic dysphonia using spectral moments of the long-term average spectrum. *Laryngoscope*. 2010;120:749–757.
36. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *J Acoust Soc Am*. 2000;108:1867–1876.
37. 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;11:74–80.
38. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103–115.
39. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice*. 2009;23:341–352.
40. Awan SN. *The Voice Diagnostic Protocol: A Practical Guide to the Diagnosis of the Voice Disorders*. Gaithersburg: Asper Publishers; 2001.
41. Awan SN, Roy N, Jette 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.
42. Peterson EA, Roy N, Awan SN, et al. Toward validation of the cepstral spectral index of dysphonia (CSID). *J Voice*. 2013;27:401–410.