Validation and Test-Retest Reliability of Acoustic Voice Quality Index Version 02.06 in the Turkish Language

[*]Gamze Yeşilli-Puzella, [†]Elçin Tadıhan-Özkan, and [‡,§,‖,¶,**]Youri Maryn, *†Eskisehir, TURKEY, and ‡Wilrijk, §‖Ghent, ¶Louvain-La-Neuve, and **Lokeren, Belgium

Summary: Objectives. The aim of the present study was to investigate the validity (both concurrent and diagnostic) and test-retest reliability of Acoustic Voice Quality Index (AVQI) version 2 (AVQI 02.06) in Turkish speaking population.

Materials and Methods. Two hundred and fifty five native Turkish subjects with normal voices (n = 128) and with voice disorders (n = 127) were asked to sustain the vowel [a:] and read aloud the Turkish phonetically balanced text. To determine the test-retest reliability of AVQI, 20 dysphonic (ie, around 15% of the group), and 20 normophonic (ie, around 15% of the group) were reassessed 15 minutes after the first AVQI determination. A three middle seconds of sustained vowel [a:] and a sentence with 25 syllables was concatenated, and AVQI analysis was conducted. The auditory-perceptual evaluation was performed by five experienced raters with Grade (G) from GRBAS Protocol.

Results. There was a statistically significant correlation between AVQI scores and auditory-perceptual evaluation of overall voice quality ($r_s = 0.717$, $P < 0.001$). AVQI gave a threshold of 2.98 for the dysphonic voice. The values of intraclass correlation coefficient with two-way mixed-effects model, single-measures type, absolute agreement definition showed an excellent test-retest reliability for AVQI in Turkish language (intraclass correlation coefficient = 0.986).

Conclusion. AVQI v.02.06 is a valid and robust tool in differentiating dysphonic and normal voice, and has excellent test-retest reliability in Turkish language.

Key Words: Acoustic voice analysis−Acoustic voice quality index−Dysphonia-voice assessment−Turkish language.

INTRODUCTION

Patients with voice problems are generally assessed by a multidisciplinary voice team consisting of at least an otolaryngologist and a speech-language pathologist (SLP).[1] The European Laryngological Society has established a multidimensional protocol for the evaluation of pathological voices: perceptual voice evaluation, videostroboscopy, acoustics, aerodynamics, and self-evaluation of the patient.[2] Auditory-perceptual evaluation is one of the most commonly used methods in the clinical environment because of its simplicity, minimal cost, and patient comfort advantages.[3] Hirano's GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale[4] and Consensus Auditory-Perceptual Evaluation of Voice[5] are the two most commonly used auditory-perceptual rating scales. However, validity and reliability of auditory-perceptual evaluation methods are generally commented due to factors related to its subjective nature (eg, listener experience), type of rating scale, and type of voice sample (eg, sustained vowel vs continuous speech).[3,6] Therefore, acoustic voice analysis methods can be applied to gain more objective results, and their noninvasiveness, low cost, and ease of use make them especially interesting in clinical voice practice.[7,8] To objectively quantify overall voice quality (ie, overall dysphonia severity) in combination of sustained vowel and continuous speech recordings, Maryn et al[8] developed the Acoustic Voice Quality Index (AVQI). Based on a statistical linear regression model, AVQI incorporates six acoustic voice parameters: smoothed cepstral peak prominence, harmonics-to-noise ratio, shimmer local, shimmer local dB, slope of the spectrum, and tilt of the regression line through the spectrum.[8] In clinical studies of dysphonia, using both sustained vowel and continuous speech is recommended to have a better ecological validity[9,10] because continuous speech represents more usual voice characteristics of the speaker and vowels are not under the influence of individual changes such as articulation, rate, and intonation.[7] AVQI operates in the software Praat.[11] It yields a single score ranging from 0 to 10. Higher scores of AVQI correspond with more dysphonic (ie, worse) voice quality.[8,12] The diagnostic accuracy and concurrent validity of AVQI as well as its acceptable responsiveness to change after voice therapy have been proven by several studies[8,13,14,15] and across different languages and phonetic contents.[16] Currently, AVQI has three different versions. AVQI version 1 (AVQIv1) was developed first[8,13] and necessitated the combined use of the freeware programs Speech Tool[17] and Praat[11] for the analysis

Accepted for publication August 14, 2020.

From the *Department of Speech and Language Therapy; Anadolu University Faculty of Health Sciences, Eskisehir, TURKEY; †Department of Speech and Language Therapy; Anadolu University Faculty of Health Sciences, Eskisehir, TURKEY; ‡European Institute for ORL-HNS, Department of Otorhinolaryngology and Head & Neck Surgery, GZA Sint-Augustinus, Wilrijk, Belgium; §Department of Rehabilitation Sciences, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium; ‖Faculty of Education, Health & Social Work, University College Ghent, Ghent, Belgium; ¶School of Logopedics, Faculty of Psychology and Educational Sciences, Université Catholique de Louvain, Louvain-La-Neuve, Belgium; and the **Phonanium, Lokeren, Belgium.

Address correspondence and reprint requests to Gamze Yeşilli-Puzella, Anadolu University, Department of Speech and Language Pathology, Yunus Emre Campus 26470 Eskisehir, Turkey. E-mail addresses: gamzeyesilli@gmail.com etadihan@anadolu.edu.tr youri@phonanium.com

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2020.08.021

of quality of voice. Subsequently, AVQI's version 2 (AVQIv2) was developed to run in only Praat. Although AVQI has been developed initially to measure dysphonia severity of Dutch speakers,[18] it has been validated in many languages from very different linguistic families. AVQIv1 and AVQIv2 have been validated for German,[16,19] French,[16] English,[16] Australian English,[20] Korean,[21,22] Finnish,[23,24] Lithuanian,[25] Japanese,[26] and Kannada.[27] The time length of sustained vowel in these two versions was 3 seconds and voiced segments in continuous speech part were less than 3 seconds. To increase its ecological validity, AVQI version 3 (AVQIv3) was developed to equalize both sustained vowel and voiced continuous speech parts to 3 seconds.[28] Barsties et al[29] evaluated the influence of gender and age on AVQI, and found that AVQI scores do not depend on gender and age. The research on test-retest reliability of AVQI showed no statistically significant differences between repeated AVQI determinations.[30]

Owing to the continuous speech part of AVQI, varying phonetic structures across different languages may have an influence on the methodological characteristics of AVQI. AVQIv2 has been investigated for different language family groups including Germanic (Dutch,[8,13,16] German,[16,19] and English[16,20]), Romance (French[16]), Baltic (Lithuanian[25]), Uralic (Finnish[23,24]), Dravidian (Kannada[27]), Altaic (Korean[21,22]), and Japonic (Japanese[26]) languages. However, as far as we know it has never been validated in Turkish. The Turkish language belongs to the Altaic language family, but different from Korean as it resides under the Turkic language subgroup (Western Altaic). Turkish has eight vowels (/ʌ /, /ɛ/, /ɯ/, /i/, /o/, /ú/, /u/, /y/) and 21 consonants (/p/, /t/, /k/, /tʃ/, /f/, /s/, /ʃ/, /h/, /b/, /d/, /g/, /dʒ/, /v/, /z/, /ʒ/, /m/, /n/, /l/, /ɾ/, /j/,[31] and soft g which is phonetically realized as the lengthening of the preceding vowel[32]), but it does not contain commonly used allophones such as [c], [ɟ], [ɫ], and [ŋ].[33] Turkish thus differs from the other languages in which AVQI has been validated. For the AVQI to be applied as a robust/valid tool for the assessment of voice quality/dysphonia in Turkish, its validity has to be assessed in that language. Because the second version of the AVQI is internationally and to date the most validated, it was decided to start investigating validity (both concurrent and diagnostic) and test-retest reliability of AVQI version 2 (AVQI 02.06) in Turkish speaking and clinically representative population.

METHODS

Subjects

This study included 255 subjects consisting of 127 dysphonic and 128 normophonic subjects, all native Turkish speakers. The dysphonic subjects had different organic and nonorganic laryngeal pathologies. They were examined at different otolaryngology departments of university hospitals and private clinics in İstanbul and Eskişehir. As a part of clinical evaluation, whispered voice hearing test was used for detecting a possible hearing impairment because the data in the literature supports that it is a simple and accurate screening for detecting hearing impairment.[34] Dysphonic participants who passed this whispered voice hearing screening were included, and those who failed the test were excluded and referred for an in depth evaluation by an audiologist. There were 72 women and 55 men, with ages ranging from 8 to 78 years (mean = 36.95 years, standard deviation (SD) = 18.53). Based on the laryngeal pathologies and the voice quality disturbances, this sample was regarded to be representative for a clinical voice disorders population in Turkey. Demographic data and laryngological diagnoses of these dysphonic subjects are presented in Table 1.

The normophonic group was composed of volunteers who reported no voice complaints, history of chronic laryngeal diseases or speech and hearing problems. These subjects consisted of patient relatives and staff and students working or studying at the universities where the voices were recorded. Since they did not report or seek help for any voice, speech or hearing problem, they were only assessed with the voice handicap index (VHI),[35,36] and this did not show any relevant impact of voice on quality of life (mean VHI = 1.02, SD VHI = 1.33). They were free from any kind of upper airway infections and hearing problems at the time of recording. This group consisted of 75 women and 53 men with ages ranging in age from 8 to 74 years (mean = 35.60 years, SD = 18.48 years.

Voice samples

The voice samples of the present study were taken in Uskudar University (Istanbul, Turkey), Yeditepe University (Istanbul, Turkey), and Anadolu University DILKOM Speech and Language Disorders Center (Eskisehir, Turkey). All the recordings were conducted in acoustically isolated lab rooms with the ambient noise level below 40 dB (first room), 50 dB (second room), and 30 dB (third room) (a-weighted level). All the recordings had a signal-to-noise ratio (SNR) above 30 dB, with mean SNR = 35.11 dB (SD = 5.9 dB), minimum SNR = 30.01 dB, and maximum SNR = 51.22 dB. All signals were thus recorded in accordance with the standard as suggested by Deliyski et al.[37,38]

The voice samples were recorded using a Shure SM48 microphone (Shure Inc., Niles, Illinois) and the Computerized Speech Lab model 4500 (KayPENTAX, Lincoln Park, New Jersey) at both Uskudar University and Anadolu University. At Yeditepe University, a cardioid AT2005 USB microphone (Audio-Technica, Stow, Ohio) was used together with Audacity software. All voices were sampled at 44.1 kHz and 16 bits of resolution.

The microphone was positioned at 10 cm from the mouth and at an azimuthal angle of circa 90°, which is in accordance to Titze & Winholtz (1993).[39] The subjects, while seated, were asked to sustain the vowel [a:] during more than 3 seconds and then to read a text at a comfortable pitch and loudness. For the continuous speech part, the third sentence of the "Pinokyo Passage" was used. This Pinokyo Passage is a phonetically balanced text, which traditionally has been used in Turkish voice clinics. This sentence of 25

TABLE 1.
Demographic Data and Laryngeal Diagnosis of Dysphonic Group

Diagnosis	Total Number (n =127)	%	Age (Years)		F (n = 72) %56.69	M (n = 55) %43.30
			Mean	SD		
Nodules	33	25.9	22.63	13.65	20	13
Sulcus vocalis	16	12.5	39.62	12.67	6	10
Unilateral paralysis/paresis	14	11.0	47.71	16.01	11	3
Functional dysphonia	12	9.4	38.91	16.97	8	4
Cyst	8	6.2	33	19.48	7	1
Mutational falsetto	7	5.5	21.57	7.69	0	7
Presbylarynx	7	5.5	67.57	7.89	3	4
Spasmodic dysphonia	4	3.1	51.75	14.24	1	3
Vocal fold scar	5	3.9	44.6	15.51	5	0
Polyp	4	3.1	38.25	9.77	1	3
Polypoid mucosa (edema)	4	3.1	48.5	15.86	4	0
Reflux laryngitis	3	2.3	34	16.64	2	1
Chronic hyperplastic laryngitis	2	1.5	48.5	26.16	0	2
Acute laryngitis	2	1.5	28	8.48	2	0
Postphonosurgery	1	0.7	56	-	0	1
Parkinson disease	1	0.7	52	-	0	1
ALS	1	0.7	45	-	1	0
Juvenile voice	1	0.7	22	-	1	0
Larynx trauma	1	0.7	17	-	0	1
Postcordectomy	1	0.7	65	-	0	1

Abbreviations: ALS, Amyotrophic lateral sclerosis; F, female; M, male.

syllables used for the continuous speech task is written as "Serüvenim resimde gördüğünüz doğa harikası şu dağ köyünde başladı." and is pronounced as [sɛɾyʋɛnim ɾɛsimde ɟúɾdyːnyz doːʌ hʌːɾikʌsɯ ʃu dʌː cújyndɛ bʌʃɫʌdɯ].

Ethics

This study was approved by the Ethics Committee of Anadolu University, Eskişehir, Turkey (28.06.2018/68159). Each participant or parent/caregiver in case of participants under 18 years old were asked to sign a consent form prior to the voice recording.

Auditory-perceptual ratings

The voice recordings were saved in wav format, and *Praat*[11] software was used to concatenate the three middle seconds of sustained vowel, followed by 2 seconds of pause and then the continuous speech part for the auditory-perceptual evaluation as described by Maryn et al.[8]

For auditory-perceptual ratings, four SLPs and one otolaryngologist with mean experience of 11 years on voice evaluation, diagnosis, and therapy rated all the voice samples. All the raters listened to the concatenated voice samples presented in random order, and judged the overall voice quality on G (Grade) scale of the GRBAS protocol[4] with a four-point Likert scale (0 = normal voice quality, 1 = slight dysphonia, 2 = moderate dysphonia, and 3 = severe dysphonia). The raters were blinded to diagnosis, demographic data of the subjects, as well as each other's auditory-perceptual ratings.

The judges received the voice recordings, anchor voices, and rating instructions electronically and they were allowed to listen to the voice samples without restriction in time and with as much repetitions as needed using around-ear headphones. To augment the jury's intrarater and inter-rater reliability, eight anchor voices were presented.[40] The anchor voices were chosen by three experienced SLPs with a minimum experience of 15 years on voice disorders from an existing voice data bank. Per G score (G = 0, G = 1, G = 2, and G = 3), one male and one female voice were included. The raters were instructed to listen to the anchor voices before the start of the experimental rating sessions, after every twenty-fifth score, and any time they wished to calibrate their perception's sensitivity.

To assess the intrarater reliability, 26 randomly chosen voice samples (13 dysphonic and 13 normophonic, ie, 10% of the total number of voice samples) were presented a second time at the end of the experiment.

Test-retest reliability

To determine the test-retest reliability of AVQIv2, 20 dysphonic (ie, around 15% of the group), and 20 normophonic (ie, around 15% of the group) were reassessed 15 minutes after the first AVQI determination.

Acoustic measures

For acoustic analysis, AVQI script version 02.06 developed by Maryn et al[8] (available at www.phonanium.com) was applied with *Praat* [11] version 6.1.02 (available at www.

praat.org). The middle 3 seconds of sustained vowel [a:] and the voiced segments of the continuous speech were analyzed with the following regression formula: AVQI = 9.072−0.245 × CPPS−0.161 × HNR−0.470 × SL + 6.158 × SLdB−0.071 × Slope−0.170 × Tilt.

Statistics

All the statistical analyses of this study were conducted using *SPSS* for Windows version 20.0 (IBM Corp., Armonk, New York). The results were considered statistically significant at $P < 0.05$. The inter-rater reliability of five raters for the G scores was assessed using the intraclass correlation coefficient (ICC): two-way random-effects model, average-measures type, and absolute agreement definition.[41] The intrarater reliability of the five raters for the 26 repeated voice samples was also examined with the ICC: two-way mixed-effects model, single-measures type, and absolute agreement definition.[41]

To investigate the criterion-related concurrent validity of AVQI v.02.06 for Turkish language, the Spearman rank-order correlation coefficient (r_s) between G_{mean} and AVQI as well as the coefficient of determination (r_s^2) was calculated. Interpretation guidelines were provided by Frey et al.[42]

The receiver operating characteristics (ROC) curve analysis was applied to evaluate the power of AVQIv2 in distinguishing between dysphonic and normophonic voices. As mentioned by Barsties and Maryn,[28] AVQI's specificity and sensitivity may change according to the AVQI threshold. In the present study the absence of hoarseness was defined as G_{mean} from 0.0 to 0.2. This means that hoarseness was decided to be present as soon as more than one of the five raters judged a sample to be slightly hoarse (or worse). A voice was considered to be normal only when four of the five raters agreed on its normality, and a voice was considered dysphonic if more than one rater scored it slightly dysphonic. The best threshold level for AVQIv2 for Turkish was calculated using Youden index. Youden index was used to evaluate the overall discriminative power of diagnostic procedure of AVQI-TR, and was calculated with the formula: (sensitivity+specificity)-1.

Spearman rank-order correlation coefficient (r_s) was determined to assess the proportional relationship between test and retest scores of AVQIv2. Kolmogorov-Smirnov test indicated that the two data sets were not normally distributed (first set: test statistic = 0.200, $P < 0.001$; second set: test statistic = 0.213, $P < 0.001$). Therefore, nonparametric Wilcoxon signed rank test was used to determine whether there is a difference between test and retest of AVQIv2. ICC with two-way mixed-effects model, single-measures type and absolute agreement definition[41] was used to calculate the test-retest reliability of AVQIv2.

RESULTS

Inter-rater reliability

Table 2 demonstrates the inter-rater reliability of the G scores of the five raters. The ICC values ranged from 0.786 (between rater one and rater five) to 0.967 (between rater three and rater four), with mean inter-rater ICC = 0.799. This indicates good inter-rater reliability.[43]

TABLE 2.
Inter-Rater Reliability of Five Raters for G Scores

	Rater 2	Rater 3	Rater 4	Rater 5
Rater 1	0.793	0.809	0.81	0.786
Rater 2		0.936	0.936	0.931
Rater 3			0.967	0.906
Rater 4				0.907

All with P <0.05.

TABLE 3.
Intrarater Reliability of Five Raters for G Scores

Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
0.970	0.967	0.973	0.984	0.957

All with P <0.001.

Intrarater reliability

Table 3 demonstrates the intrarater reliability of the G scores of the five raters. The ICC values of intrarater reliability ranged from 0.957 (rater five) to 0.984 (rater four) with a mean value of ICC = 0.970. This reflects excellent intrarater reliability.[43]

Criterion-related concurrent validity

The r_S between AVQIv2 scores and G_{mean} was 0.717 ($P <$ 0.001) signifying an acceptable relationship and concurrent validity. This result indicates that higher AVQI scores are related to higher ratings of G. Figure 1 illustrates the proportional relationship between G_{mean} and AVQIv2 for Turkish. The $r_S^2 = 0.514$ indicated however that only half (51.4%) of the variance in G_{mean} is accounted for by the variance in AVQIv2.

Diagnostic accuracy

In order to determine the best cut-off value to differentiate between dysphonic and normal voices, ROC curve was computed (Figure 2). Sensitivity, specificity and area under ROC curve (A_{ROC}) values were considered to observe the accuracy of AVQIv2 for Turkish. $G_{MEAN} = 0.2$ was used as an indicator of dysphonia by reason of the idea that if more than one of the five raters evaluated a voice to be slightly dysphonic, that voice would have an abnormal quality. With $A_{roc} = 0.896$, AVQIv2 showed acceptable diagnostic accuracy for the Turkish population. The Youden index computed an optimum cut-off value of 2.98 with a sensitivity of 0.747 (74.7%) and specificity of 0.923 (92.3%).

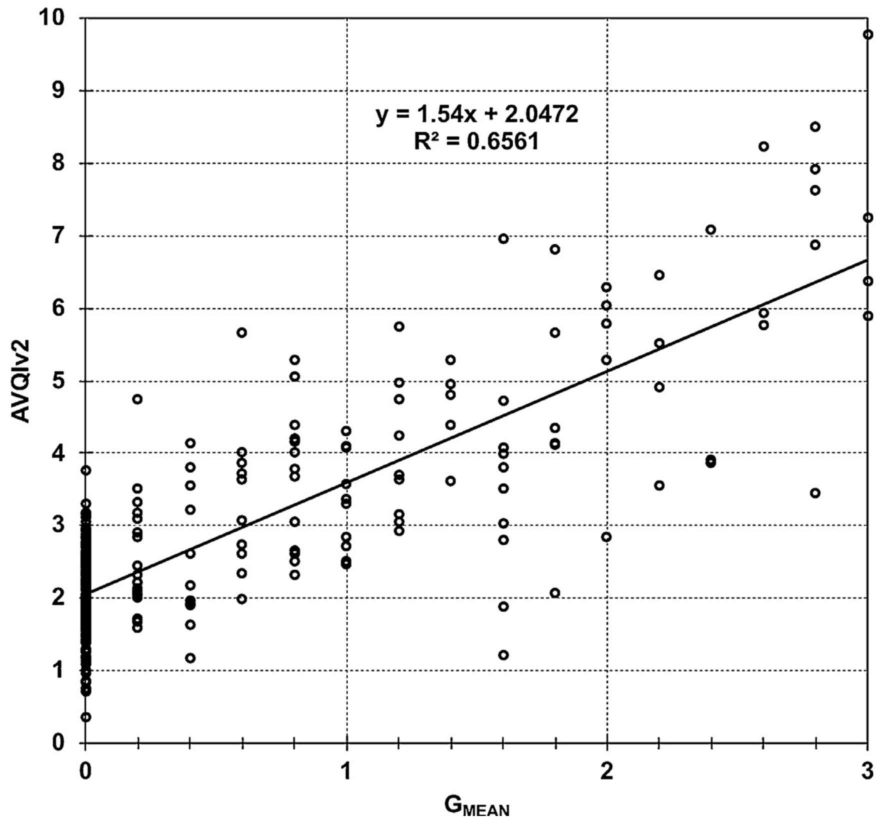

FIGURE 1. Scatterplot with linear regression fit line between G_{MEAN} and AVQIv2.

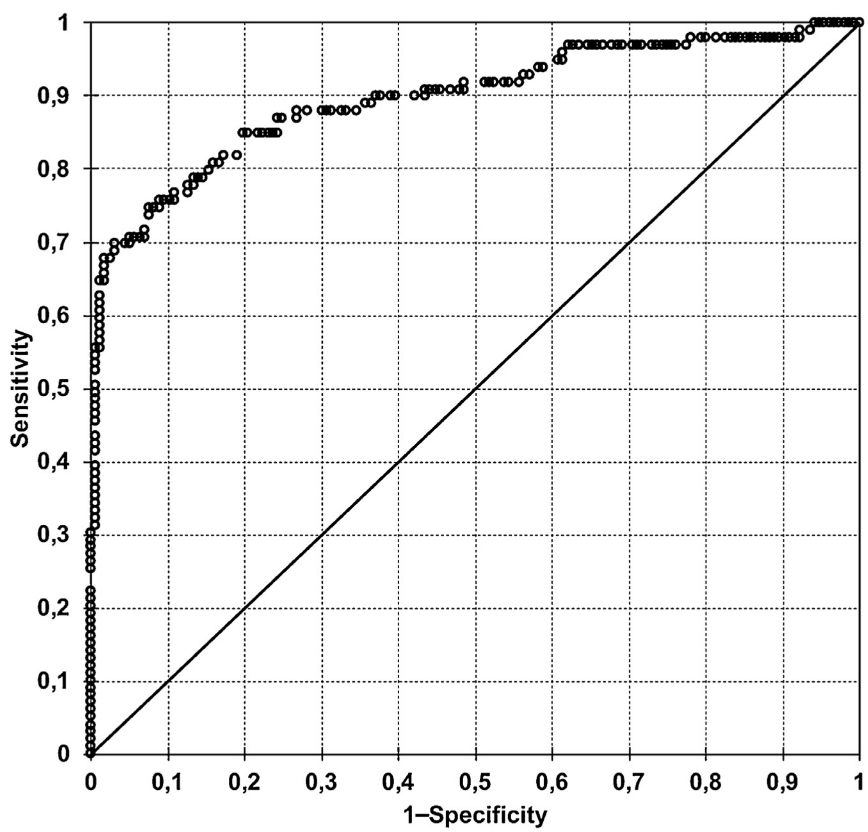

FIGURE 2. Receiver operating characteristics curve illustrating the diagnostic accuracy of AVQIv2 in Turkish language.

Test-retest reliability

With $r_s = 0.909$ ($P < 0.001$), Spearman correlation established an almost perfect correlation between test and retest scores of AVQIv2. With $Z = -1.015$ ($P = 0.31$), Wilcoxon signed rank test revealed no differences between test and retest scores of AVQIv2. With ICC = 0.986, AVQIv2 showed an excellent test-retest reliability.[43]

DISCUSSION

The present study investigated the criterion-related concurrent validity, the diagnostic accuracy, and the test-retest reliability of AVQIv2 (currently AVQI v.02.06) for the Turkish language. Before doing this, however, acceptable inter- and intrarater reliability of auditory-perceptual assessment was ensured.

The criterion-related concurrent validity of AVQIv2 was investigated based on auditory-perceptual judgment. In this study, the r_S between AVQIv2 scores and G_{mean} was 0.717 ($P < 0.001$), signifying an acceptable relationship and concurrent validity.

This result of concurrent validity of AVQIv2 is similar to that found in Dutch,[8,13] English,[20] French,[16] and Finnish[24] languages (ie, $r_s = 0.78$,[8] $r_s = 0.796$,[13] $r_s = 0.794$,[20] $r_s = 0.781$,[16] and $r_s = 0.75$,[24] respectively) with the perceptual ratings based on G parameter from GRBAS protocol. This result is also similar to that found in German[19] language ($r_s = 0.79$) which is based on 'H' (Hoarseness) from RBH (roughness, breathiness, and hoarseness) scale.[44] Since hoarseness from RBH is equal to G from the GRBAS, and they are both measured based on a four-point ordinal scale,[45] it is possible to compare the results of the two studies. Other studies across different languages showed higher concurrent validity between AVQI and G parameter, values ranging from $r_s = 0.808$ to $r_s = 0.911$.[16,28,21,25,26,22]

The accuracy of AVQIv2 in differentiating between Turkish dysphonic and normophonic voices was also acceptable, as expressed in $A_{ROC} = 0.896$. This result is similar to Dutch[8,16] ($A_{ROC} = 0.895$ and $A_{ROC} = 0.893$), and German[19] ($A_{ROC} = 0.888$). The highest values were found for Dutch[13,28] ($A_{ROC} = 0.920$ and $A_{ROC} = 0.94$), German[16] ($A_{ROC} = 0.958$), English[16] ($A_{ROC} = 0.953$), Lithuanian[25] ($A_{ROC} = 0.940$), and Korean[22] ($A_{ROC} = 0.975$) languages. The diagnostic threshold identified in the present study is 2.98, with 74.7 % sensitivity and 92.3 % specificity. This is similar to the threshold values for Dutch[13] (2.95), Lithuanian[25] (2.97), and Finnish[24] (2.87) languages. Other languages found threshold values ranging from 2.36 to 3.66.[8,16,19,20,22,26,28] The sensitivity and specificity of this study is similar to the those of German[19] (ie, sensitivity = 79 %, specificity = 92 %), and Japanese[26] (ie, sensitivity = 72.5 %, specificity = 95.2 %). The sensitivity level found in this study is lower than in other studies with different languages ranging from 79.6 % to 98 %.[8,13,16,20,22,24,25,28] The result of this study indicates that AVQIv2 for Turkish language has a moderate sensitivity to correctly identify the dysphonic subjects, and an excellent specificity to correctly identify the normal voices. We observed discrepancy between sensitivities and specificities across the many studies so far. However, in this kind of investigation, there are multiple factors that may cause differences between the outcomes of the studies next to language and phonetic structure of the sentences (especially in terms of voiced percentage). First, on what basis are study subjects allocated in normophonia versus dysphonia groups: complaints, dysphonia severity rating and its arbitrarily chosen threshold, and/or laryngeal imaging? In the present study, mean G below/above 0.2 served as threshold, but there are other AVQI validation studies that have applied other thresholds, as for example Maryn et al[8] who used G = 0 as threshold. Second, what is the level of experience of those providing the auditory-perceptual ratings? Auditory perception of voice quality relies to some degree on experience.[46,47,48] In this study there were five raters with 11 years on average of clinical experience, whereas in other studies this kind of information is not available.[8,16,21,22,24,25,26] So currently it is not possible to give additional substantiation for this thought. Third, what type of rating scale (ordinal as in GRBAS, or continuous as in Consensus Auditory-Perceptual Evaluation of Voice) was used? Fourth, culture-based differences of perception of dysphonia may play a role. The differences in the perception of health, illnesses, and disabilities across cultures have been proved by some researches[49,50] and similarly, the effect of the cultural and language backgrounds of the judges on the perception of some voice quality types has been proved,[51] and it was also found that the perception of roughness is language dependent.[52] As far as we know there is no sole explanation for this. To summarize, AVQI for Turkish language revealed a moderate sensitivity to correctly identify the dysphonic subjects, which is comparable to that of German[19] and Japanese.[26]

In the present study, test-retest reliability of AVQIv2 for Turkish language was investigated. The results revealed that AVQIv2 has an excellent test-retest reliability, as indicated by an ICC = 0.986. This result is similar to that found by Barsties and Maryn[30] (ie, ICC = 0.97).

CONCLUSION

The AVQI is an easy-to-use, and quick instrumental tool for the objective measure of overall voice quality. It is possible to say that AVQI for Turkish language can be easily used for clinical and research purposes, not only for population of 83.154 million living in Turkey, but also Turkish speakers living abroad. The results of this study revealed that AVQI v.02.06 is a valid and robust tool in differentiating dysphonic and normal voice, and has excellent test-retest reliability in Turkish language. Future research of AVQI for the assessment of pediatric voice disorders for Turkish language would be recommended.

REFERENCES

1. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-based clinical voice assessment: a systematic review. *Am J Speech-Language Pathol*. 2013;22:212–226. https://doi.org/10.1044/1058-0360(2012/12-0014).
2. 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: guideline elaborated by the committee on phoniatrics of the European laryngolo. *Eur Arch Oto-Rhino-Laryngology*. 2001;258:77–82. https://doi.org/10.1007/s004050000299.
3. Oates J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatr Logop*. 2009;61:49–56. https://doi.org/10.1159/000200768.
4. Hirano M. Psycho-acoustic evaluation of voice. In: Arnold G, Winckel F, Wyke B, eds. *Disorders of Human Communication 5. Clinical Examination of Voice*. Vienna, Austria: Springer-Verlag; 1981:81−84.
5. Kempster GB, Gerratt BR, Abbott KV, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech-Language Pathol*. 2009. https://doi.org/10.1044/1058-0360(2008/08-0017).
6. Awan SN, Roy N, Zhang D, et al. Validation of the Cepstral Spectral Index of Dysphonia (CSID) as a screening tool for voice disorders: development of clinical cutoff scores. *J Voice*. 2016;30:130–144. https://doi.org/10.1016/j.jvoice.2015.04.009.
7. Parsa V, Jamieson DG. Acoustic discrimination of pathological voice. *J Speech, Lang Hear Res*. 2001;44:327–339. https://doi.org/10.1044/1092-4388(2001/027).
8. Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010;24:540–555. https://doi.org/10.1016/j.jvoice.2008.12.014.
9. Lu FL, Matteson S. Speech tasks and interrater reliability in perceptual voice evaluation. *J Voice*. 2014;28:725–732. https://doi.org/10.1016/j.jvoice.2014.01.018.
10. 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. https://doi.org/10.1016/j.jvoice.2004.08.009.
11. Boersma P, Praat. a system for doing phonetics by computer. *Glot Int*. 2002.
12. Batthyany C, Maryn Y, Trauwaen I, et al. A case of specificity: how does the acoustic voice quality index perform in normophonic subjects? *Appl Sci*. 2019;9:1–12. https://doi.org/10.3390/app9122527.
13. Maryn Y, De Bodt M, Roy N. The acoustic voice quality index: toward improved treatment outcomes assessment in voice disorders. *J Commun Disord*. 2010;43:161–174. https://doi.org/10.1016/j.jcomdis.2009.12.004.
14. Barsties B, Maryn Y. External validation of the acoustic voice quality index version 03.01 with extended representativity. *Ann Otol Rhinol Laryngol*. 2016;125:571–583. https://doi.org/10.1177/0003489416636131.
15. Uloza V, Latoszek BB v, Ulozaite-Staniene N, et al. comparison of dysphonia severity index and acoustic voice quality index measures in differentiating normal and dysphonic voices. *Eur Arch Oto-Rhino-Laryngology*. 2018;275:949–958. https://doi.org/10.1007/s00405-018-4903-x.
16. Maryn Y, De Bodt M, Barsties B, et al. The value of the acoustic voice quality index as a measure of dysphonia severity in subjects speaking different languages. *Eur Arch Oto-Rhino-Laryngology*. 2014;271:1609–1619. https://doi.org/10.1007/s00405-013-2730-7.
17. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech, Lang Hear Res*. 1996. https://doi.org/10.1044/jshr.3902.311.
18. Maryn Y, Weenink D. Objective dysphonia measures in the program praat: Smoothed cepstral peak prominence and acoustic voice quality index. *J Voice*. 2015;29:35–43. https://doi.org/10.1016/j.jvoice.2014.06.015.
19. Barsties B, Maryn Y. Der acoustic voice quality index in Deutsch. *HNO*. 2012;60:715–720. https://doi.org/10.1007/s00106-012-2499-9.
20. Reynolds V, Buckland A, Bailey J, et al. Objective assessment of pediatric voice disorders with the acoustic voice quality index. *J Voice*. 2012;26. https://doi.org/10.1016/j.jvoice.2012.02.002. 672.e1-672.e7.
21. Maryn Y, Kim HT, Kim J. Auditory-perceptual and acoustic methods in measuring dysphonia severity of Korean speech. *J Voice*. 2016;30:587–594. https://doi.org/10.1016/j.jvoice.2015.06.011.
22. Kim GH, Lee YW, Bae IH, et al. Validation of the acoustic voice quality index in the Korean language. *J Voice*. 2018;33:948.e1–948.e9. https://doi.org/10.1016/j.jvoice.2018.06.007.
23. Kankare E, Barsties B, Maryn Y, et al. A preliminary study of the acoustic voice quality index in Finnish speaking population. A Prelim study. *Acoust Voice Qual Index Finnish Speak Popul*. 2015;218. (August).
24. Kankare E, Barsties V, Latoszek B, et al. The acoustic voice quality index version 02.02 in the Finnish-speaking population. *Logop Phoniatr Vocol*. 2019;0:1–8. https://doi.org/10.1080/14015439.2018.1556332.
25. Uloza V, Petrauskas T, Padervinskis E, et al. Validation of the acoustic voice quality index in the Lithuanian language. *J Voice*. 2017;31:257.e1–257.e11. https://doi.org/10.1016/j.jvoice.2016.06.002.
26. Hosokawa K, Barsties B, Iwahashi T, et al. Validation of the acoustic voice quality index in the Japanese language. *J Voice*. 2017;31:260.e1–260.e9. https://doi.org/10.1016/j.jvoice.2016.05.010.
27. Kishore Pebbili G, Shabnam S, Pushpavathi M, et al. Diagnostic accuracy of acoustic voice quality index version 02.03 in discriminating across the perceptual degrees of dysphonia severity in Kannada language. *J Voice*. 2019. https://doi.org/10.1016/j.jvoice.2019.07.010.
28. Barsties B, Maryn Y. The improvement of internal consistency of the acoustic voice quality index. *Am J Otolaryngol Head Neck Med Surg*. 2015;36:647–656. https://doi.org/10.1016/j.amjoto.2015.04.012.
29. Barsties V, Latoszek B, Ulozaite-Staniene N, et al. The Influence of gender and age on the acoustic voice quality Index and dysphonia severity index: a normative study. *J Voice*. 2017:38–40. https://doi.org/10.1016/j.jvoice.2017.11.011.
30. Barsties B, Maryn Y. Test-retest-variabilität und interne konsistenz des acoustic voice quality index. *HNO*. 2013;61:399–403. https://doi.org/10.1007/s00106-012-2649-0.
31. Topbaş S, Yavaş M.Communication Disorders in Turkish.; 2010. doi:10.3109/02699206.2010.526286.
32. Ünal-Logacev Ö, Zygis M, Fuchs S. Phonetics and phonology of soft "g" in Turkish. *J Int Phon Assoc*. 2019;49:183–206. https://doi.org/10.1017/S0025100317000317.
33. Göksel A, Kerslake C. *Turkish: A Comprehensive Grammar*. London: Routledge; 2005:3–9. https://doi.org/10.4324/9780203340769.
34. Pirozzo S, Papinczak T, Glasziou P. Whispered voice test for screening for hearing impairment in adults and children: systematic review. *Br Med J*. 2003;327:967–970. https://doi.org/10.1136/bmj.327.7421.967.
35. Jacobson BH, Johnson A, Grywalski C, et al. The Voice Handicap Index (VHI): development and validation. *Am J Speech-Language Pathol*. 1997. https://doi.org/10.1044/1058-0360.0603.66.
36. Kiliç MA, Okur E, Yildirim İ, et al. Ses Handikap Endeksi (Voice Handicap Index) Türkçe versiyonunun güvenilirliği ve geçerliliği. *Kulak Burun Boğaz İhtisas Dergisi*. 2008;18(3):139–147.
37. Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. *J Voice*. 2005;19:15–28. https://doi.org/10.1016/j.jvoice.2004.07.003.
38. Deliyski DD, Shaw HS, Evans MK, et al. Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatr Logop*. 2006;58:274–288. https://doi.org/10.1159/000093184.
39. Titze IR, Winholtz WS. Effect of microphone type and placement on voice perturbation measurements. *J Speech Hear Res*. 1993. https://doi.org/10.1044/jshr.3606.1177.
40. Rossi M, Pavlovic CV, Espesser R. Reducing context effects in the subjective evaluation of speech quality. *J Acoust Soc Am*. 1990. https://doi.org/10.1121/1.2028481.
41. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016. https://doi.org/10.1016/j.jcm.2016.02.012.

42. Frey L, Botan C, Kreps G. Investigating communication. *NY Allyn Bacon*. 2000. http://uurimismeetodid.pbworks.com/f/00+Investigating+Communication.+Contents+and+Preface.PDF.
43. Portney LG, Watkins MP. *Foundations of clinical research: applications to practice*. 2000.
44. Nawka T, Anders LC, Wendler J, et al. Die auditive Beurteilung heiserer Stimmen nach dem RBH-System. 1994.
45. Barsties B, De Bodt M. Assessment of voice quality: current state-of-the-art. *Auris Nasus Larynx*. 2015;42:183–188. https://doi.org/10.1016/j.anl.2014.11.001.
46. 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. https://doi.org/10.1016/S0892-1997(97)80026-4.
47. Wolfe VI, Martin DP, Palmer CI. Perception of dysphonic voice quality by naive listeners. *J Speech, Lang Hear Res*. 2000. https://doi.org/10.1044/jslhr.4303.697.
48. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006. https://doi.org/10.1016/j.jvoice.2005.08.007.
49. Hampton NZ, Marshall A. Culture, gender, self-efficacy, and life satisfaction: a comparison between Americans and Chinese people with spinal cord injuries. *J Rehabil*. 2000;66:21.
50. Diener E, Diener M. Cross-cultural correlates of life satisfaction and self-esteem. *Culture and Well-Being*. Springer; 2009:71–91.
51. 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. https://doi.org/10.1159/000119746.
52. Ghio A, Cantarella G, Weisz F, et al. Is the perception of dysphonia severity language-dependent? A comparison of French and Italian voice assessments. *Logop Phoniatr Vocology*. 2015;40:36–43. https://doi.org/10.3109/14015439.2013.837503.