Development and Validation of the Japanese Version of the Consensus Auditory-Perceptual Evaluation of Voice

AUTHOR(S):

Kondo, Kanako; Mizuta, Masanobu; Kawai, Yoshitaka; Sogami, Tohru; Fujimura, Shintaro; Kojima, Tsuyoshi; Abe, Chika; ... Tateya, Ichiro; Omori, Koichi; Haji, Tomoyuki

CITATION:

Kondo, Kanako ...[et al]. Development and Validation of the Japanese Version of the Consensus Auditory-Perceptual Evaluation of Voice. Journal of Speech, Language, and Hearing Research 2021, 64(12): 4754-4761

ISSUE DATE:

2021-12-13

URL:

http://hdl.handle.net/2433/275629

Development and validation of the Japanese version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Short title: Development of the Japanese CAPE-V

Kanako Kondo, SLP[1]

Masanobu Mizuta, MD, PhD[1,2]

Yoshitaka Kawai, MD, PhD[1]

Tohru Sogami, MD[3]

Shintaro Fujimura, MD, PhD[1]

Tsuyoshi Kojima, MD, PhD[4]

Chika Abe, SLP[2]

Ryo Tanaka, SLP[5]

Osamu Shiromoto, PhD[6]

Ryuji Uozumi, PhD[7]

Yo Kishimoto, MD, PhD[1]

Ichiro Tateya, MD, PhD[8]

Koichi Omori, MD, PhD[1]

Tomoyuki Haji, MD, PhD[2]

1. Department of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan
2. Department of Otolaryngology, Head and Neck Surgery, Kurashiki Central Hospital, Kurashiki, Japan
3. Department of Otorhinolaryngology, Soseikai General Hospital, Kyoto, Japan
4. Department of Otolaryngology, Tenri Hospital, Nara, Japan

5. Rehabilitation center, Tenri Hospital, Nara, Japan
6. Department of Communication Sciences and Disorders, Faculty of Health and Welfare, Prefectural University of Hiroshima, Mihara, Japan
7. Department of Biomedical Statistics and Bioinformatics, Graduate School of Medicine, Kyoto University, Kyoto, Japan
8. Department of Otolaryngology - Head and Neck Surgery, School of Medicine, Fujita Health University, Toyoake, Japan

✓ The authors report no conflicts of interest.

Address correspondence to:

Yo Kishimoto, MD, PhD; Department of Otolaryngology-Head & Neck Surgery, Graduate School of Medicine; Kyoto University, Sakyo-ku, Kyoto 606-8507, Japan. Tel: +81-75-751-3346, Fax: +81-75-751-7225, E-mail: y_kishimoto@ent.kuhp.kyoto-u.ac.jp

ABSTRACT

Purpose: Auditory perceptual evaluation is essential for the assessment of voice quality. The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) provides a standardized protocol and assessment form for clinicians to analyze the voice quality and has been adapted into several different languages. The aims of this study were to develop the Japanese version of the CAPE-V and to investigate its reliability and validity.

Method: The Japanese CAPE-V consisted of the same three speech contexts (vowels, sentences, and conversation) as developed in the original English version. The sentences were designed according to the concepts of the original version and reviewed by Japanese phoneticians. To validate the usefulness of the Japanese CAPE-V, voices of 173 Japanese-speaking subjects (76 subjects with dysphonia and 97 without voice complaints) were evaluated by 5 experienced judges, according to the Japanese CAPE-V as well as the GRBAS scale.

Results: The Japanese CAPE-V provided a high inter-rater reliability (intraclass correlation coefficients [ICC] > 0.85 for all the parameters) as well as a high intra-rater reliability (ICC > 0.85 for all the parameters). In addition, overall severity, roughness, and breathiness in the Japanese CAPE-V were highly correlated with the corresponding dimensions in the GRBAS scale, having Spearman's correlation coefficients greater than 0.8.

Conclusions: The current study demonstrated the reliability and validity of the newly developed Japanese CAPE-V as an auditory perceptual evaluation instrument.

Key Words: Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), Japanese language, auditory perceptual analysis

INTRODUCTION

To evaluate the quality of one's voice, it is necessary to conduct multidimensional assessments, including auditory perceptual evaluation, subjective assessment, acoustic analysis, and aerodynamic analysis. Since voice is fundamentally perceptual, auditory perceptual evaluation of voice has greater intuitive meaning and shared reality among listeners than instrumental measures do (Oates, 2009). In addition, most patients seek treatment for voice disorders based on a perceived disruption in voice quality (Carding et al., 2009). Consequently, auditory perceptual evaluation is often considered the gold standard for assessment of voice disorders (Awan et al., 2016; Oates, 2009). Additionally, auditory perceptual evaluation can be performed noninvasively and does not require electrical equipment, which enhances its usefulness in the clinic.

In general, an assessment system of any health status, including voice quality, should be reliable and valid for it to be clinically used. Reliability is the degree to which an instrument is free from random error, and validity is the degree to which the instrument measures what it purports to measure (Aaronson et al., 2002). Auditory perceptual evaluation is subjective and thus can be influenced by the rater's experience and rating systems, which leads to variability and errors (Oates, 2009; Wuyts et al., 1999). Therefore, the development of a well-designed and controlled rating system is necessary to enhance the usability of auditory perceptual evaluation (Oates, 2009).

To date, many auditory perceptual evaluation systems have been proposed, with the goal of having a standardized protocol that would allow for sharing of information among clinicians (Carding et al., 2009; Hammarberg et al., 1980; Hirano, 1981). The GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale was developed by the Japanese Phoniatrica Committee (Hirano, 1981) and consists of 5 perceptual dimensions on a four-point ordinal scale (0: normal, 1: mild, 2: moderate, 3: severe) for each dimension. Since this method is easy to use and allows clinicians to quickly evaluate a voice disorder, it has been utilized universally. In Japan, the GRBAS method is the most common evaluation system, and sustained vowels are the recommended voice samples to be analyzed with the GRBAS scale (Phoniatrics, 2009). In addition, high reliability and validity of the GRBAS scale have been reported (De Bodt et al., 1997; Webb et al., 2004; Wuyts et al., 1999). However, the GRBAS scale has several limitations. Although the four-point scale of this method allows raters to evaluate easily, it is difficult to detect slight differences in dysphonia severity. In addition, the GRBAS scale using sustained vowels lacks ecological validity. Most patients with voice disorders typically complain of connected speech during conversation and not of problems with sustained vowels. Therefore, the GRBAS scale using sustained vowels may not evaluate the actual, everyday voice quality of patients.

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed by the American Speech-Language-Hearing Association (ASHA) to provide a standardized protocol and assessment form for clinicians to analyze the voice quality of patients (Kempster et al., 2009). This tool contains three phonatory tasks: sustained vowels (/a/, /i/), reading six sentences, and conversational speech. The CAPE-V employs visual analog scales (a 100-mm line) for judging the six parameters, including overall severity, roughness, breathiness, strain, pitch, and loudness. That visual analog scale includes the following three textual markers underneath the scale: MI = mildly deviant, MO = moderately deviant, SE = severely deviant. In 2009, Kempster et al. published a peer-reviewed article including a reprint of the CAPE-V form. In the published version, the textual markers were distributed nonlinearly, while in the revised version, the markers were placed linearly at 10, 50, and 90 mm (Nagle et al., 2014). Consequently both the "nonlinearly-distributed" and "linearly-distributed version" versions of the CAPE-V form are in use (Nagle, 2016). Previous studies reported the high reliability and validity of the CAPE-V (Karnell et al., 2007; Kempster et al., 2009; Zraick et al., 2011).

In addition, the CAPE-V has a strong relationship to acoustic estimates for both continuous speech and vowels from the CAPE-V's task using cepstral/spectral acoustic measurements (Awan et al., 2010; Awan et al. 2013; Mizuta et al., 2020; Watts, 2015).

As a result of the usefulness of the original CAPE-V, it has been translated into different languages: Brazilian Portuguese (Behlau et al., 2020; Nemr et al., 2016), European Portuguese (de Almeida et al., 2019), Italian (Mozzanica et al., 2013), Spanish (Núñez-Batalla et al., 2015), Turkish (Chen et al., 2018; Ertan-Schlüter et al., 2020; Özcebe et al., 2019), and Mandarin (Chen et al., 2018). The reliability and validity of the adaptations of the CAPE-V were investigated and demonstrated in each language (Behlau et al., 2020; Chen et al., 2018; de Almeida et al., 2019; Ertan-Schlüter et al., 2020; Mozzanica et al., 2013; Nemr et al., 2016; Núñez-Batalla et al., 2015; Özcebe et al., 2019). Thus, the CAPE-V has been utilized around the world. However, there has been no Japanese language version of the CAPE-V developed to date. The purposes of this study were to adapt the CAPE-V to the Japanese language and evaluate its reliability and the validity.

METHOD

Development of the Japanese version of the CAPE-V

Prior to starting this study, permission was obtained from ASHA to develop the Japanese version of the CAPE-V. The Japanese CAPE-V consisted of the same three tasks as the original English version. Task 1 was performed in the same way as the original version. Subjects were asked to utter the vowels /a/ and /i/ for 3-5 seconds. In Task 2, six specific sentences were designed to elicit different laryngeal behaviors. These sentences were adapted to Japanese contexts according to the concepts of the original version and reviewed by Japanese phoneticians (Table 1). Sentence 1 was designed to study the coarticulatory influence of various vowels. The adapted sentence for the Japanese version was "Yoru no bento wa Kuri-gohan da" (I will get takeout Kuri-gohan, a steamed rice with chestnut, for my dinner). Sentence 2 provoked the production of soft glottic attacks and transitions from voiceless to voiced. The adapted sentence was "Haha wa hana ni hohoemu" (Mother smiles when she sees flowers). Sentence 3 produced all the voiced phonemes and created appropriate context for judging the possible existence of spasms/detentions and the subject's ability to link one word to another. The adapted sentence was "Wara no yane no ie da" (That house has a straw roof). Sentence 4 included many words that began with a vowel and could elicit hard glottic attacks, providing the opportunity to see whether these occurred in the subject. The adapted sentence was "Ima, ikki ni ita wo kiru" (I just cut the board). Sentence 5 included numerous nasal consonants, providing the chance to evaluate hyponasality and assess whether stimulation by resonant voice therapy was possible. The adapted sentence was "Nandemo mama no mane dane" (You always copy your mother's habit). Sentence 6 had no nasal consonants and created a useful context for evaluating intraoral pressure and possible hypernasality and nasal air emission. The adapted sentence was "Pirrito karai kakinotane wo katta" (I bought a pack of Kakinotane, the hot and spicy rice crackers with peanuts). Task 3 was running speech. The clinician interviewed the subjects and allowed them to talk naturally for at least 20 seconds. Topics were about the subject's voice and preference, for example, "Tell me about your voice problem," "Tell me about your favorite foods/music," and "What are you interested in?".

Subjects

Native Japanese-speaking people participated in this study. The subjects comprised 76 patients with dysphonia and 97 volunteers without voice complaints. The dysphonic patients were recruited from Kyoto University Hospital, Kurashiki Central Hospital, or Tenri Hospital, while the volunteers without voice complaints were recruited from the employees of those institutions. The subject characteristics are shown in Table 2. All procedures in this study conformed to the Declaration of Helsinki and were approved by the Institutional Review Board of each institution. All participants provided written informed consent.

All recordings were made in an acoustically treated room, with noise below 50 dB. All voices were captured with a condenser microphone (audio-technica AT4050, Shure SM35-XLR, AKG C747), placed 4 cm away from the mouth. Voice samples were recorded in WAV format with a 44.1 kHz sampling rate and 32-bit quantization.

Rating

After the voice samples were randomized, auditory perceptual evaluation was performed by 5 specialists (3 otolaryngologists and 2 speech language pathologists) in a quiet room to eliminate background noises. Each specialist judge had over 5 years of experience in evaluating voices on the GRBAS scale, although they had never used a visual analog scale for auditory perceptual evaluation. The judges listened to the voice samples with headphones (AKG, K240MK II) on a Windows personal computer through a free application software for sound reproduction (foobar2000, https://www.foobar2000.org/) to maintain a consistent volume.

Before the rating session, each judge practiced on 10 sample recordings. After listening to all the voice tasks of each subject, the judges rated the subject's performance according to the original CAPE-V protocol. When the subject's performance was not uniform across tasks, their overall performance was rated. The rating was calculated based on the CAPE-V, which employs a visual analog scale on 6 perceptual dimensions: overall severity, roughness, breathiness, strain, pitch, and loudness. The visual analog scale used in this study was the published version of the CAPE-V (Kempster et al., 2009), which has a 100-mm line with linearly distributed severity labels for mild, moderate, and severe at approximately 10, 50, and 90 mm along the line. Data were collected on paper. In addition, the voice samples were evaluated based on the GRBAS scale, which consists of 5 perceptual dimensions (grade, roughness, breathiness, asthenia, and strain) on a four-point ordinal scale (0: normal, 1: mild, 2: moderate, 3: severe). After evaluating 20 voices, the judges took a 10-minute break. The CAPE-V rating and the GRBAS rating were performed with at least a 1-week interval, and the evaluations could be made in any sequence. In addition, for intra-rater reliability analysis of the CAPE-V, all voice samples were re-rated based on the CAPE-V after at least a 1-week interval.

In order to examine the validity between the CAPE-V and the GRBAS scales , the CAPE-V or the GRBAS rating was performed for all voice samples with at least 1-week interval. In addition, more than 1 week later, all voice samples were re-rated based on the CAPE-V for intra-rater reliability analysis

Statistical Analysis

The intra-rater reliability and inter-rater reliability were assessed using the intraclass correlation coefficients (ICC) (Shrout et al., 1979) computed based on a random effects model with average measures. To determine the concurrent validity between the CAPE-V and the GRBAS scales on 4 corresponding perceptual dimensions, Spearman's correlation coefficients were calculated. The mean rating of the five judges was used as the reference value for the correlation analysis. Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc. Cary, NC).

RESULTS

Reliability analysis

Intra-rater reliability was examined by calculating the ICC for each of the CAPE-V parameters. As shown in Table 3, the ICC values for all raters were higher than 0.85 for all parameters. The highest ICC value was 0.98 (95% confidence interval (CI), 0.97-0.98) for overall severity, whereas the lowest ICC value was 0.86 (95% CI, 0.83-0.89) for pitch.

The inter-rater reliability results across all raters are shown in Table 4. As with intra-rater reliability, the ICC values of inter-rater reliability were higher than 0.85 for all the parameters. The highest ICC value was 0.97 (95% CI, 0.93-0.98) for overall severity, while pitch had the lowest ICC value of 0.86 (95% CI, 0.81-0.89).

Validity analysis

Figure 1 shows the relationship between the CAPE-V and the GRBAS scale ratings on the 4 corresponding perceptual dimensions. To determine the concurrent validity between the CAPE-V and the GRBAS scales, Spearman's correlation coefficients were calculated, as shown in Table 5. Overall severity, roughness, and breathiness in the CAPE-V were highly correlated with the corresponding dimensions in the GRBAS scale, having correlation coefficients of over 0.8. The correlation of strain between the two scales was low, with a correlation coefficient of 0.67 (95% CI, 0.58-0.75).

DISCUSSION

Auditory perceptual evaluation is essential for the assessment of voice quality. In this study, the Japanese version of the CAPE-V was developed, and the reliability and the validity were investigated. All the sentences for the speech tasks in the Japanese CAPE-V were designed to fulfill the phonetic targets and purposes established by the original CAPE-V and were reviewed by Japanese phoneticians.

The results in this study showed a strong agreement among the 5 judges in both the intra-rater and the inter-rater reliability. In particular, the ICC values for overall severity, roughness, breathiness, strain, and loudness were higher than 0.9. Previous studies in different languages of the CAPE-V showed consistent results (Table 6) (de Almeida et al., 2019; Ertan-Schlüter et al., 2020; Mozzanica et al., 2013; Zraick et al., 2011; Özcebe et al., 2019), demonstrating cross-lingual robustness of the CAPE-V. In particular, the inter-rater reliability for overall severity was the highest among the 6 perceptual dimensions in all the previous studies described in Table 6, which is consistent with the current result. On the other hand, the reliability for pitch was lower than that of other parameters in the current study, which is not consistent with previous studies (Table 6) (de Almeida et al., 2019; Ertan-Schlüter et al., 2020; Mozzanica et al., 2013; Zraick et al., 2011; Özcebe et al., 2019). This might be due to insufficient training of the judges for voice assessment. In Japan, the GRBAS scale has been the most commonly used auditory perceptual evaluation system, and thus, the judges in this study were well-trained in auditory perceptual analysis for overall severity, roughness, breathiness, and strain, parameters the GRBAS scale measures. In contrast, as pitch is not included in the GRBAS scale, the training provided in this study might not have been enough for them to assess pitch, which might have led to the lower ICC value of this parameter.

The validity of the Japanese CAPE-V was investigated through calculating correlation coefficients between the four comparable CAPE-V and GRBAS parameters. Overall severity, roughness, and breathiness presented strong correlations between the two scales, whereas strain showed a moderate correlation (Schober et al., 2018). The GRBAS scale is widely used, being the most common instrument in Japan, and its validity has been demonstrated in previous studies (De Bodt et al., 1997; Webb et al., 2004; Wuyts et al., 1999). The moderate-to-strong correlation of the GRBAS scale with the Japanese CAPE-V protocol in the current study suggests that our new protocol is valid. As with the current reliability study described above, this result of the validity is consistent with that of previous studies of the CAPE-V in different languages (de Almeida et al., 2019; Karnell et al., 2007; Mozzanica et al., 2013; Zraick et al., 2011). Notably, the correlation coefficient between the CAPE-V and the GRBAS rating was the strongest in overall severity (Grade) for all studies described in Table 6.

There are several advantages of the CAPE-V over the GRBAS scale. The GRBAS scale provides no specific protocol for universal administration, although sustained vowels are recommended as the voice samples to analyze in Japan. The CAPE-V defines an elicitation protocol and uses ecologically valid conversational speech probes, which include phonetically diverse speech contexts (Zraick et al., 2011), justifying the CAPE-V protocol. In addition, a hybrid visual analog scale is employed in the CAPE-V, whereas a categorical scale with four points is utilized in the GRBAS scale. The visual analog scale, which generates continuous data, allows the use of parametric analysis and provides greater statistical power, enabling the CAPE-V to detect smaller differences in voice quality within and across subjects compared to the GRBAS scale. On the other hand, the GRBAS scale is easier to use in clinical settings because a four-point ordinal scale for each dimension is employed. Clinicians can choose the adequate auditory perceptual instrument based on their interests and capabilities.

There are several limitations that warrant future studies. First, the current study employed the published version of the CAPE-V with linearly distributed severity labels. On the other hand, it was demonstrated that raters tend to exhibit a systematic bias at the lower end of interval scales when equally distributed interval scaling is utilized (1975 Stevens, 2016 Nagle). Therefore, the revised version of the CAPE-V employed nonlinearly distributed labels (2014 Nagle). In future study, investigation of the Japanese version of the CAPE-V with nonlinear labels would be warranted. Second, since the voices were recorded at different institutions, three different microphones were utilized in this study, which has the potential to affect the results. In future studies, consistent microphone models should be used. Finally, all judges in this study were highly experienced in auditory perceptual evaluation. It is yet to be determined whether less experienced judges can rate at a level consistent with well-experienced judges. In addition, the learning curve of the judges is an important topic of future studies, as it is important to understand how much training is needed for consistent judgement.

CONCLUSIONS

The Japanese version of the CAPE-V was developed to promote a universally standardized approach for auditory perceptual assessment. The current study revealed that the Japanese CAPE-V provides high inter-rater and intra-rater reliabilities as well as high validity, which is demonstrated through moderate-to-strong correlations of the corresponding parameters with the GRBAS scale. These results justify the use of the Japanese CAPE-V in clinical practice.

Acknowledgments

This study was supported by a research grant from the Japan Society of Logopedics and Phoniatrics, and Grants-in-Aid for Scientific Research (KAKENHI). We thank the phoneticians, Drs. Ken-ichi Sakakibara and Michiko Hashi, and the Speech-Language Pathologist, Dr. Yoshiaki Ozawa, for review of the developed sentences in the Japanese CAPE-V.

References

Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research, 11*(3), 193-205. https://doi.org/10.1023/a:1015291021312

Awan, S. N., Roy, N., Jetté, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clinical Linguistics and Phonetics, 24*(9), 742-758. https://doi.org/10.3109/02699206.2010.492446

Awan, S. N., Roy, N., Zhang, D., & Cohen, S. M. (2016). Validation of the Cepstral Spectral Index of Dysphonia (CSID) as a Screening Tool for Voice Disorders: Development of Clinical Cutoff Scores. *Journal of Voice, 30*(2), 130-144. https://doi.org/10.1016/j.jvoice.2015.04.009

Awan, S. N., Solomon, N. P., Helou, L. B., & Stojadinovic, A. (2013). Spectral-cepstral estimation of dysphonia severity: external validation. *Annals of Otology, Rhinology, and Laryngology, 122*(1), 40-48. https://doi.org/10.1177/000348941312200108

Behlau, M., Rocha, B., Englert, M., & Madazio, G. (2020). Validation of the Brazilian Portuguese CAPE-V Instrument-Br CAPE-V for Auditory-Perceptual Analysis. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.07.007

Carding, P. N., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: state of the science review. *Journal of Laryngology and Otology, 123*(8), 823-829. https://doi.org/10.1017/S0022215109005398

Chen, Z., Fang, R., Zhang, Y., Ge, P., Zhuang, P., Chou, A., & Jiang, J. (2018). The Mandarin Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) and Its Reliability. *Journal of Speech, Language, and Hearing Research, 61*(10), 2451-2457. https://doi.org/10.1044/2018_JSLHR-S-17-0386

de Almeida, S. C., Mendes, A. P., & Kempster, G. B. (2019). The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) Psychometric Characteristics: II European Portuguese Version (II EP CAPE-V). *Journal of Voice, 33*(4), 582.e585-582.e513. https://doi.org/10.1016/j.jvoice.2018.02.013

De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Croux, C. (1997). Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11*(1), 74-80. https://doi.org/10.1016/s0892-1997(97)80026-4

Ertan-Schlüter, E., Demirhan, E., Ünsal, E. M., & Tadıhan-Özkan, E. (2020). The Turkish Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): A

Reliability and Validity Study. *Journal of Voice, 34*(6), 965.e913-965.e922. https://doi.org/10.1016/j.jvoice.2019.05.014

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-laryngologica, 90*(5-6), 441-451. https://doi.org/10.3109/00016488009131746

Hirano, M. (1981). *Clinical Examination of Voice*. New York: Springer-Verlag.

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21*(5), 576-590. https://doi.org/10.1016/j.jvoice.2006.05.001

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124-132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Mizuta, M., Abe, C., Taguchi, E., Takeue, T., Tamaki, H., & Haji, T. (2020). Validation of Cepstral Acoustic Analysis for Normal and Pathological Voice in the Japanese Language. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.08.026

Mozzanica, F., Ginocchio, D., Borghi, E., Bachmann, C., & Schindler, A. (2013). Reliability and validity of the Italian version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Folia Phoniatrica et Logopaedica, 65*(5), 257-265. https://doi.org/10.1159/000356479

Nagle, K. (2016). Emerging scientist: challenges to CAPE-V as a standard. *Perspectives of the ASHA Special Interest Groups, 1*, 47-53.

Nagle, K. F., Helou, L. B., Solomon, N. P., & Eadie, T. L. (2014). Does the presence or location of graphic markers affect untrained listeners' ratings of severity of dysphonia? *Journal of Voice, 28*(4), 469-475. https://doi.org/10.1016/j.jvoice.2013.12.011

Nemr, K., Simões-Zenari, M., de Souza, G. S., Hachiya, A., & Tsuji, D. H. (2016). Correlation of the Dysphonia Severity Index (DSI), Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), and Gender in Brazilians With and Without Voice Disorders. *Journal of Voice, 30*(6), 765.e767-765.e711. https://doi.org/10.1016/j.jvoice.2015.10.013

Núñez-Batalla, F., Morato-Galán, M., García-López, I., & Ávila-Menéndez, A. (2015). Validation of the Spanish adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Acta Otorrinolaringologica Española, 66*(5), 249-257. https://doi.org/10.1016/j.otorri.2014.07.007

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatrica et Logopaedica, 61*(1), 49-56.

https://doi.org/10.1159/000200768

Phoniatrics, t. J. S. o. L. a. (2009). *Auditory Perceptual Evaluation in Voice Assessment, new edition (in Japanese)*. Tokyo: Ishiyaku Publishers.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and Analgesia, 126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037//0033-2909.86.2.420

Watts, C. R. (2015). The Effect of CAPE-V Sentences on Cepstral/Spectral Acoustic Measures in Dysphonic Speakers. *Folia Phoniatrica et Logopaedica, 67*(1), 15-20. https://doi.org/10.1159/000371656

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology, 261*(8), 429-434. https://doi.org/10.1007/s00405-003-0707-7

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). 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. *Jounal of Voice, 13*(4), 508-517. https://doi.org/10.1016/s0892-1997(99)80006-x

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., . . . Glaze, L. E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology, 20*(1), 14-22. https://doi.org/10.1044/1058-0360(2010/09-0105)

Özcebe, E., Aydinli, F. E., Tiğrak, T. K., İncebay, Ö., & Yilmaz, T. (2019). Reliability and Validity of the Turkish Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Journal of Voice, 33*(3), 382.e381-382.e310. https://doi.org/10.1016/j.jvoice.2017.11.013

Figure legend

Figure 1

Relationships between subjects' CAPE-V and GRBAS ratings on overall severity (grade), roughness, breathiness, and strain.

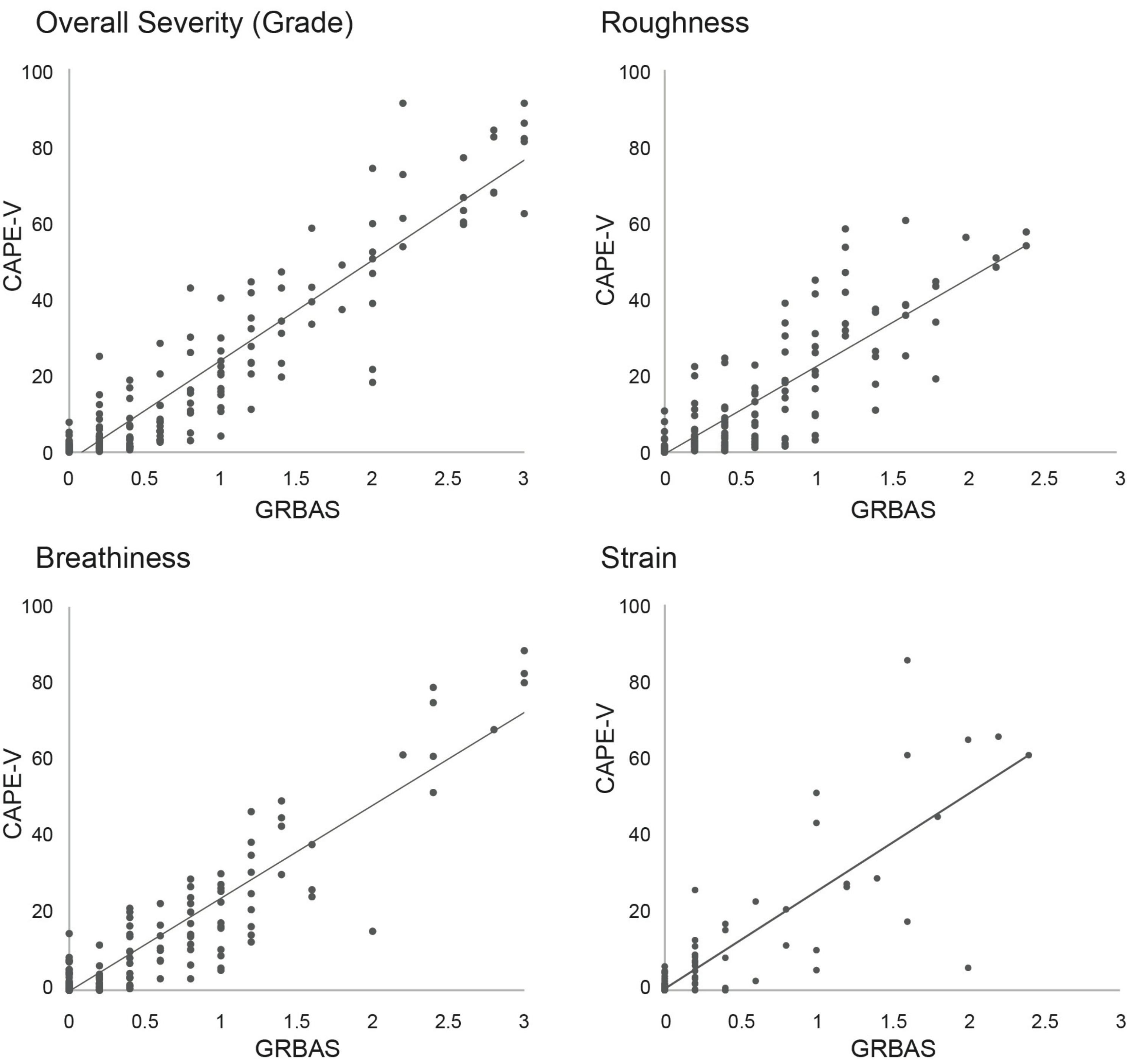
Overall Severity (Grade)
Roughness
Breathiness
Strain
CAPE-V
GRBAS
0
20
40
60
80
100
0.5
1
1.5
2
2.5
3

Table 1: Adapted sentences in the Japanese vesion of the CAPE-V

	Japanese version	IPA Transcription
Sentence 1	夜の弁当はくりごはんだ。	[joɾɯnobentoːɰa kɯɾʲigoɦanda.]
Sentence 2	母は花にほほえむ。	[haɦaɰa hanaɲihoɦoemɯ.]
Sentence 3	わらの屋根の家だ。	[ɰaɾanojanenoieda.]
Sentence 4	今、一気に板をきる。	[ima ikkʲiɲiitaokʲiɾɯ.]
Sentence 5	なんでもママのマネだね。	[nandemomamanomanedane.]
Sentence 6	ピリッと辛い柿の種を買った。	[pʲiɾɾʲitto kaɾai kakʲinoʈaneokatta.]

Original English version
The blue spot is on the key again.
How hard did he hit him.
We were away a year ago.
We eat eggs every Easter.
My mama makes lemon muffins.
Peter will keep at the peak.

Table 2. Subject Characteristics

Gender	
Male	70
Female	103
Age (year)	
Average (Standard Deviation)	46 (17)
Range	17-83
Diagnostic category	
Benign vocal fold lesion	28
Unilateral vocal fold paralysis	14
Muscle tension dysphonia	9
Spasmodic dysphonia	7
Vocal fold atrophy	4
Essential tremor	3
Scarred vocal folds	2
Acute laryngitis	2
Trauma	2
Others	5
Volunteer	97

Table 3. Intra-rater Reliability of the Japanese CAPE-V

Vocal Parameter	ICC (95% CI)
Overall severity	0.98 (0.97, 0.98)
Roughness	0.96 (0.95, 0.97)
Breathiness	0.97 (0.96, 0.97)
Strain	0.96 (0.95, 0.97)
Pitch	0.86 (0.83, 0.89)
Loudness	0.96 (0.95, 0.97)

CAPE-V: Consensus Auditory-Perceptual Evaluation of Voice , ICC: intraclass correlation coefficient, CI: confiddence interval

Table 4. Inter-rater Reliability of the Japanese CAPE-V

Vocal Parameter	ICC (95% CI)
Overall severity	0.97 (0.93, 0.98)
Roughness	0.95 (0.91, 0.97)
Breathiness	0.96 (0.93, 0.98)
Strain	0.96 (0.94, 0.97)
Pitch	0.86 (0.81, 0.89)
Loudness	0.96 (0.94, 0.97)

CAPE-V: Consensus Auditory-Perceptual Evaluation of Voice , ICC: intraclass correlation coefficient, CI: confiddence interval

Table 5. Correlation Between the CAPE-V and the GRBAS Scales

CAPE-V	GRBAS	Spearman's correlation coefficient (95% CI)
Overall severity	Grade	0.89 (0.85, 0.92)
Roughness	Roughness	0.83 (0.77, 0.87)
Breathiness	Breathiness	0.81 (0.76, 0.86)
Strain	Strain	0.67 (0.58, 0.75)

CAPE-V: Consensus Auditory-Perceptual Evaluation of Voice, CI: confiddence interval

Table6. Overview of the previous studies in different languages of the CAPE-V

	Zraick et al (2011)	Mozzanica et al (2013)	de Almeida (2019)
Language	English	Itarian	European Portuguese
Number of Voice samples	59	200	20
Number of Raters	21	3	14
Intrarater reliability statistics	ICC	ICC	Pearson correlation
Overall Severity	0.57	0.92	0.87
Roughness	0.77	0.92	0.61
Breathiness	0.82	0.90	0.87
Strain	0.35	0.89	0.73
Pitch	0.78	0.88	0.92
Loudness	0.64	0.80	0.69
Inter-rater reliability statistics	ICC	ICC	ICC
Overall Severity	0.76	0.92	0.96
Roughness	0.62	0.91	0.92
Breathiness	0.60	0.90	0.95
Strain	0.56	0.76	0.84
Pitch	0.54	0.83	0.86
Loudness	0.28	0.82	0.90
Correlation statistics parameter parameter	Multiserial correlations	Spearman correlations	Multiserial correlations
Overall Severity (Grade)	0.80	0.92	0.95
Roughness	0.76	0.84	0.89
Breathiness	0.78	0.87	0.90
Strain	0.77	0.79	0.47

ICC: intraclass correlation coefficient

Özcebe et al (2019)	Ertan-Schlüter et al (2020)	Present study
Turkish	Turkish	Japanese
130	182	173
4	2	5
ICC	Pearson correlation	ICC
0.93-0.96	0.92	0.98
0.89-0.96	0.86	0.96
0.89-0.98	0.87	0.97
0.76-0.86	0.82	0.96
0.88-0.99	0.81	0.86
0.86-0.97	0.84	0.96
ICC	ICC	ICC
0.90	0.95	0.97
0.81	0.91	0.95
0.84	0.93	0.96
0.80	0.88	0.96
0.81	0.89	0.86
0.88	0.92	0.96
Spearman correlations	Spearman correlations	Spearman correlations
0.80	0.85	0.89
0.62	0.82	0.83
0.67	0.77	0.81
0.68	0.66	0.67