CROSS-CULTURAL ADAPTATION OF THE CONSENSUS AUDITORY-PERCEPTUAL EVALUATION OF VOICE (CAPE-V) INTO MALAY: VALIDITY AND RELIABILITY STUDY

NURHAYATI BINTI MOHD MOSSADEQ

UNIVERSITI SAINS MALAYSIA

2023

CROSS-CULTURAL ADAPTATION OF THE CONSENSUS AUDITORY-PERCEPTUAL EVALUATION OF VOICE (CAPE-V) INTO MALAY: VALIDITY AND RELIABILITY STUDY

by

NURHAYATI BINTI MOHD MOSSADEQ

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

February 2023

ACKNOWLEDGEMENT

In the name of Allah, the Beneficent, the Merciful. Alhamdulillah, praise Allah for giving me strength and His blessing, helping me throughout this journey, and finally, coming to an end.

I would like to express my thanks to my main supervisor, Dr. Khairy Anuar bin Mohd Khairuddin for his excellent guidance, support, and enthusiasm. A special thanks to my co-supervisor, Associate Professor Dr. Mohd Normani bin Zakaria for all the valuable input.

Special thanks to my dedicated colleague Dr. Jamilah, How Hui Teng, Haniza, Nur Shahira, Rosmah Suriani, and Sherin for participating in the expert panels. Thank you to all participants from the Hospital Kuala Lumpur, Hospital Putrajaya, and Hospital Universiti Sains Malaysia for spending their time in data collection. I am also very thankful to all the dedicated staff of the Speech Therapy Unit, Hospital Kuala Lumpur, Speech-Language Therapy and Occupational Therapy Unit, Hospital Putrajaya, and Speech-Language Clinic, School of Health Sciences, Universiti Sains Malaysia for their helpful support.

I owe my heartfelt gratitude to my husband, Mohd Yusrin bin Saparin for his unending encouragement and prayers. Thank you to my siblings, Hanif and Nur Atiqah, for your support. To my gorgeous boys and girl, Muhammad Zawir, Muhammad Zihni, and Nurzuyyin Mardhiyah, thank you for your understanding and apologies for my frequent absence. I hope this thesis became an inspiration to all of you. Finally, this thesis is dedicated to my late parents, Mohd Mossadeq bin Ismail and Khamisah binti Asit, as well as my late brother, Fadhli. Al Fatihah.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ASHA	American Speech-Language-Hearing Association
CAPE-V	Consensus Auditory-Perceptual Evaluation of Voice
CVI	Content Validity Index
EAIS	Equal-Appearing Interval Scale
GRBAS Scale	Grade-Roughness-Breathiness-Asthenia-Strain Scale
ICC	Intraclass Correlation Coefficient
I-CVI	Item-level Content Validity Index
ORL	Otorhinolaryngology
S-CVI/Ave	Scale-Level Content Validity Index Based on the Average
SLT	Speech-Language Therapist
VAS	Visual Analogue Scale

LIST OF APPENDICES

ADAPTASI SILANG BUDAYA *CONSENSUS AUDITORY-PERCEPTUAL EVALUATION OF VOICE* (CAPE-V) KE BAHASA MELAYU: KAJIAN KESAHAN DAN KEBOLEHPERCAYAAN

ABSTRAK

Perkadaran suara auditori-perseptual merupakan salah satu penilaian suara klinikal yang perlu dilaksanakan. Salah satu alat yang mendapat perhatian untuk tujuan ini adalah *Consensus Auditory-Perceptual Evaluation of Voice* (CAPE-V). CAPE-V dilaporkan mempunyai pendekatan yang standad dalam menilai suara. Dengan kelebihan ini, CAPE-V telah diadaptasi ke beberapa bahasa lain untuk disesuaikan dengan populasi spesifik. Oleh itu, adaptasi CAPE-V ke bahasa Melayu adalah penting untuk penggunaannya di kalangan populasi di Malaysia seperti yang ingin dicapai di dalam kajian ini. Kajian keratan rentas ini melibatkan penterjemahan CAPE-V ke dalam bahasa Melayu, diikuti pengenalpastian kesahan dan kebolehpercayaannya. Proses penterjemahan (hadapan dan belakang) melibatkan empat orang penterjemah yang mempunyai pengalaman sekurang-kurangnya 14 tahun. Kebanyakan item diterjemahkan dengan sama oleh kesemua penterjemah. Item yang tidak konsisten terjemahannya ditentukan terjemahan yang paling sesuai melalui persetujuan semua penterjemah. Analisis kesahan kandungan melibatkan dua orang penilai yang mempunyai pengalaman klinikal sekurang-kurangnya 12 tahun, dengan menggunakan kaedah CVI. Nilai CVI mencapai aras boleh terima, oleh itu, kesemua item dikekalkan. Bagi analisis kesahan gagasan, kesahan serentak, kebolehpercayaan antara penilai dan kebolehpercayaan penilai-sama, CAPE-V yang telah diterjemahkan, dikenali sebagai *Malay* CAPE-V, digunakan untuk mengadarkan rakaman sampel suara daripada 38 peserta, yang terdiri daripada 19 individu dengan suara normal (6 lelaki dan 11

perempuan) dan 19 individu dengan masalah suara (8 lelaki dan 11 perempuan). Perkadaran ini dijalankan oleh tiga penilai yang mempunyai sekurang-kurangnya 12 tahun pengalaman klinikal. Kesahan gagasan telah diuji dengan menggunakan kaedah *known-group* yang melibatkan perbandingan skor setiap parameter vokal Malay CAPE-V di antara kumpulan bersuara normal dan bermasalah suara melalui analisis ujian Mann-Whitney U, di mana analisis menunjukkan terdapat perbezaan yang signifikan di antara kedua-dua kumpulan tersebut. Bagi kesahan serentak, hubungan di antara skor setiap parameter vokal Malay CAPE-V dan GRBAS *Scale* dikenalpasti dengan menggunakan ujian korelasi Spearman, yang menunjukkan perhubungan positif sangat kuat yang signifikan di antara kedua-dua alat penilaian tersebut. Bagi kebolehpercayaan antara penilai dan kebolehpercayaan penilai-sama, kesemua nilai ICC skor setiap parameter vokal Malay CAPE-V adalah lebih tinggi daripada aras boleh terima. Oleh itu, *Malay* CAPE-V merupakan alat penilaian auditori-perseptual yang sah dan boleh dipercayai dalam menilai masalah suara dalam kalangan populasi Malaysia.

CROSS-CULTURAL ADAPTATION OF THE CONSENSUS AUDITORY-PERCEPTUAL EVALUATION OF VOICE (CAPE-V) INTO MALAY: VALIDITY AND RELIABILITY STUDY

ABSTRACT

Auditory-perceptual rating of voice is one of the compulsory clinical voice evaluations. One tool that has gained attention for this purpose is the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). The CAPE-V has been reported to have a more standardized approach in evaluating the voice. Because of its advantages, the CAPE-V has been adapted in several languages to suit specific populations. Accordingly, the adaptation of the CAPE-V into Malay is essential for its utility among the Malaysian population, which this study aims to achieve. This cross-sectional study involves translating the CAPE-V into Malay, followed by establishing the validity and reliability of the translated CAPE-V. The translation processes (i.e., forward and backward) involved 4 translators with at least 14 years of experience. Most items were similarly translated by the translators, except for a few, which were subsequently accepted as similar, following consensus among the translators. The content validity analysis involved 2 raters, with at least 12 years of clinical experience, using the CVI method. The CVI value was above the acceptable level, thus, all items were retained. For the analyses of construct validity, concurrent validity, inter-rater reliability, and intra-rater reliability, the translated CAPE-V, termed the Malay CAPE-V, was used to rate the voice samples recorded from 38 participants, which consisted of 19 individuals with normal voices (6 males and 13 females) and 19 individuals with voice disorders (8 males and 11 females). The ratings were performed by 3 raters with at least 12 years of clinical experience. The construct validity was examined by using

the known-group method involving a comparison of the scores of each vocal parameter of the Malay CAPE-V of the normal and disordered voice groups using the Mann-Whitney U test, which indicated a statistically significant difference between the two groups. For the concurrent validity, the relationship between the scores of each vocal parameter of the Malay CAPE-V and GRBAS Scale were examined using the Spearman correlation, which indicated a statistically significant very strong positive correlation between the two tools. For the inter-rater reliability and intra-rater reliability, the ICC values of the scores of each vocal parameter of the Malay CAPE-V were all above the acceptable level. Therefore, the Malay CAPE-V is a valid and reliable tool for an auditory-perceptual rating of voice among the Malaysian population.

CHAPTER 1

INTRODUCTION

1.1 Cross-Cultural Adaptation of a Tool

Cross-cultural adaptation of a tool is a systematic modification of an evidence-based tool to make it compatible with the individual's cultural patterns, meaning, and values by taking into consideration the language used, as well as the culture and context in which it is being used (Castro et al., 2010; Gjersing et al., 2010). This procedure is essential since the tool will be utilized in a different language, environment, and time (Gjersing et al., 2010) than the original tool. While it is compulsory for the tool to be translated linguistically, the protocols, however, must be preserved as similar to the original version as possible (Beaton et al., 2000). Only after the required process has been completed, the adapted tool can be used for the targeted population.

1.2 Auditory-Perceptual Rating of Voice

The auditory-perceptual rating of voice is one of the routines in voice assessment, in which the clinician listens to the voice samples to judge the normality or abnormality of the voice (Carding et al., 2009; Zraick et al., 2011). Several tools for the auditory-perceptual rating of voice have been introduced, such as the Grade-Roughness-Breathiness-Asthenia-Strain (GRBAS) Scale (Hirano, 1981), Vocal Profile Analysis Scheme (Laver et al., 1981), and Buffalo Voice Profile (Wilson, 1987). Among these tools, the GRBAS Scale (Appendix A) is arguably the most commonly used in clinical practice (Behlau et al., 2020; Mcalister et al., 2020; Nagle, 2016; Nemr et al., 2016; Zraick et al., 2011). While it is valid as a tool for the perceptual evaluation of voice, the lack of a conclusive protocol and interpretation

guideline may affect the utility of the GRBAS Scale (Kempster et al., 2009; Kreiman et al., 2007, 1993; Zraick et al., 2011), which then points to the need for a more standardized tool for the auditory-perceptual rating of voice.

Realizing the standardization issue of the GRBAS Scale, the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed (Kempster et al., 2009) (Appendix B). The CAPE-V is a clinician-administered tool in which the speech-language therapist (SLT) listens to the recorded samples of vocal tasks performed by the patient to judge the deficiency in the vocal parameters, if any. The CAPE-V was reported to have a more standardized approach in terms of administration and interpretation procedures (Kempster et al., 2009; Nagle, 2016). Additionally, with the possibility of judging more vocal parameters, the CAPE-V offers a more comprehensive evaluation of voice characteristics (Kempster et al., 2009). Further, the use of a visual analog scale (VAS), as the basis for measurement, improves the sensitivity of the CAPE-V in judging voice changes (Nemr et al., 2012).

Because of its clear advantages in clinical voice evaluation, the original CAPE-V has been adapted from English into different languages to suit the linguistic and cultural backgrounds of the particular population. The CAPE-V has been adapted into Brazilian Portuguese (Behlau et al., 2020; Nemr et al., 2012, 2016), Italian (Mozzanica et al., 2014), Spanish (Núñez-Batalla et al., 2015), European Portuguese (De Almeida et al., 2016), Turkish (Ertan-Schlüter et al., 2020; Özcebe et al., 2017), Mandarin (Chen et al., 2018), Hindi (Isha Baheti, 2019), and Kannada (Gunjawate et al., 2020).

1.3 Malay Adaptation of the CAPE-V

The CAPE-V has been translated into Malay for clinical purposes (Mohd Khairuddin, n.d.). The translation of the CAPE-V into Malay was done by an SLT with

a specialization in voice disorders by including the inputs of a linguist. For a full clinical utility, the available Malay version of the CAPE-V has to be systematically adapted, involving proper translation, and establishment of validity and reliability (Guillemin, 1995).

1.3.1 Translation

A proper translation process requires forward translation and backward translation of all content in the tool (Sousa and Rojjanasrirat, 2011; Toury, 1995). The translation process emphasizes conceptual translation instead of literal translation (Chen et al., 2018; De Almeida et al., 2016; World Health Organization, n.d.). In addition, the translation process usually requires the involvement of translators from different relevant backgrounds (Sousa and Rojjanasrirat, 2011). Specific to the sentence production task, instead of semantic-based translation, the sentences have to be developed according to the phonetic contexts of the original CAPE-V (Chen et al., 2018; De Almeida et al., 2016; Ertan-Schlüter et al., 2020).

1.3.2 Validity

Validity refers to the degree to which the tool evaluates what it aims to evaluate (Aaronson et al., 2002; Devon et al., 2007; Souza et al., 2017). Although there are several types of validity, only content, construct, and concurrent validities are usually examined (Aaronson et al., 2002; De Almeida et al., 2016; Ertan-Schlüter et al., 2020). Content validity indicates that the domain of the tool is appropriate for its intended use (Polit, 2015; Souza et al., 2017). The process requires input from a panel of professionals to judge whether the translated tool is suitable for the targeted population (Aaronson et al., 2002; Souza et al., 2017). Previous studies on the CAPE-V incorporated inputs from professionals such as SLTs (i.e., Chen et al., 2018; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Núñez-Batalla et al., 2015; Özcebe

et al., 2017), linguists (i.e., De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Jesus, Barney, Couto, et al., 2009), speech and hearing scientists (i.e., Jesus, Barney, Santos, et al., 2009), and phoneticians (i.e., Chen et al., 2018). Meanwhile, construct validity is the extent to which a set of variables accurately reflects the construct being examined (Souza et al., 2017). To fulfill this, all the relevant studies (i.e., De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Özcebe et al., 2017) used the known-group method by comparing the scores of each vocal parameter of the CAPE-V of individuals with normal and disordered voices, in which the scores of the disordered voices must be higher than those of the normal voices. The last common validity analysis is concurrent validity, which shows the extent to which the scores of the tool are related to a criterion measure (Aaronson et al., 2002; Souza et al., 2017). To show this type of validity, previous studies (i.e., Behlau et al., 2020; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Isha Baheti, 2019; Zraick, Birdwell, et al., 2005) examined the relationship between the scores of the CAPE-V and the GRBAS Scale to demonstrate that an acceptable relationship between the two tools was established.

1.3.3 Reliability

Reliability is defined as the internal consistency and stability of the tools, free from random error or unwanted variation (Aaronson et al., 2002; Carding et al., 2009; Souza et al., 2017). The common types of reliability are inter-rater reliability and intra-rater reliability (Behlau et al., 2020; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Mozzanica et al., 2014; Zraick et al., 2011). Inter-rater reliability evaluates the equivalence of ratings of a tool when used by different raters (Kimberlin and Winterstein, 2008; Polit, 2015). Meanwhile, intra-rater reliability analyzes the consistency of the same rater's rating of the same content across two or more trials (Devon et al., 2007; Polit, 2015). In both types of reliability, the tool must achieve an

acceptable value to show that inter-reliability and intra-reliability have been established.

1.4 Problem Statement

The CAPE-V is a valuable auditory-perceptual assessment tool to be used by practicing SLTs in clinical settings. The adaptation of the CAPE-V into Malay is essential to ensure an accurate voice disorder diagnosis for the Malaysian population. For a proper adaptation, the CAPE-V needs to undergo a systematic process. The first process is that the CAPE-V needs to be translated into Malay. Note that the translated CAPE-V will be termed the Malay CAPE-V henceforth. Then, the Malay CAPE-V has to undergo the examination of validity (i.e., content validity, construct validity, and concurrent validity) and reliability (i.e., inter-rater reliability and intra-rater reliability).

1.5 Study Objective

1.5.1 General objective

To adapt the CAPE-V, in English, into Malay.

1.5.2 Specific objectives

i. To translate the CAPE-V from English into Malay.

ii. To determine the validity of the Malay CAPE-V in terms of

a. content validity, by analyzing the relevancy of Malay CAPE-V items.

b. construct validity, by comparing the scores of each vocal parameter of the Malay CAPE-V of the normal and disordered voice groups.

c. concurrent validity, by determining the relationship between the scores of each vocal parameter of the Malay CAPE-V and GRBAS Scale.

iii. To determine the reliability of the Malay CAPE-V in terms of

a. inter-rater reliability, by analyzing the agreement of the scores of each vocal parameter of the Malay CAPE-V of different raters.

b. intra-rater reliability, by analyzing the agreement of the scores of each vocal parameter of the Malay CAPE-V of the same rater from different rating sessions.

1.6 Hypothesis

In relation to the objectives, there are two hypotheses for this study.

1.6.1 Hypothesis 1: Construct validity

H_O: There is no significant difference in the scores of each vocal parameter of the Malay CAPE-V between the normal and disordered voice groups.

H_A: There is a significant difference in the scores of each vocal parameter of the Malay CAPE-V between the normal and disordered voice groups.

1.6.2 Hypothesis 2: Concurrent validity

H_O: There is no significant correlation in the scores of each vocal parameter of the Malay CAPE-V and GRBAS Scale.

H_A: There is a significant correlation in the scores of each vocal parameter of the Malay CAPE-V and GRBAS Scale.

CHAPTER 2

LITERATURE REVIEW

2.1 Cross-cultural Adaptation

Cross-cultural adaptation is a systematic process of modifying a tool to be used in a new setting with a different language, time, and location (Castro et al., 2010; Gjersing et al., 2010). In this process, it is important to maintain the concept of the tool, but at the same time to consider the language and culture of the new setting (Beaton et al., 2000; Guillemin, 1995). Commonly, cross-cultural adaptation involves the translation of the original tool to the intended language, which is then followed by examining the validity and reliability of the translated tool (Guillemin, 1995).

2.1.1 Translation

Translation of a tool aims to achieve a conceptually equivalent translated version of the tool to the original version (Aaronson et al., 2002; Koller et al., 2007; Sousa and Rojjanasrirat, 2011; World Health Organization, n.d.). The recommended translation process should include forward translation and backward translation (Maneesriwongul and Dixon, 2004; Sousa and Rojjanasrirat, 2011).

In forward translation, a minimum of two translators from different relevant backgrounds, who are well-versed in the original language and translated language, are required (Beaton et al., 2000; Sousa and Rojjanasrirat, 2011). In the composition of the translators, at least each one of them must be familiar with the content of the tool and knowledgeable about the cultural and linguistic nuances of the translated language (Beaton et al., 2000; Sousa and Rojjanasrirat, 2011; World Health Organization, n.d.). For the backward translation, the translators must be different from those in the forward translation (Beaton et al., 2000; Gjersing et al., 2010; Sousa and

Rojjanasrirat, 2011). However, similar to those the forward translation, a minimum of two translators are required for the backward translation (Sousa and Rojjanasrirat, 2011). They must be fluent in the original language and translated language, with at least one of them must be familiar with the content of the original tool (Sousa and Rojjanasrirat, 2011).

The procedures of the translation start with the forward translation, which is then followed by the backward translation. The forward translation involves the translation of the items in the tool from the original language to the intended language (Beaton et al., 2000; Sousa and Rojjanasrirat, 2011; Sperber, 2004), which needs to be completed separately by the translators. If there is any dissimilarity in the translated items, a consensus between the translators should be sought to come out with the version to be used for the backward translation (Beaton et al., 2000). In the backward translation, the translated items are independently translated back to the original language without referring to the original tool (Beaton et al., 2000; Sousa and Rojjanasrirat, 2011). The backward-translated items from each translator are then compared to the original items. Any discrepancy between items from the backward translation and the original tool, as well as between the translators, should be resolved through consensus (Gjersing et al., 2010). These processes produce the translated version of the tool that would be used for the subsequent validity and reliability analyses.

2.1.1(a) Content validity

Content validity is the evidence to which the items in the tool are represented and relevant to the domain of the items (Almanasreh et al., 2019; Souza et al., 2017; Waltz et al., 2010; Yusoff, 2019). In the context of adapting a tool, the adapted tool should convey the concept and content equivalent to the original tool (Guillemin,

1995). Among the approach to analyze content validity, the content validity index (CVI) is more widely used (Almanasreh et al., 2019; Polit and Beck, 2006; Polit et al., 2007) due to its straightforwardness and simplicity (Polit et al., 2007).

To determine the CVI, a minimum number of two raters is needed to evaluate how much the items represent the content domain (Davis, 1992; Thorn and Deitz, 1989; Yusoff, 2019). The CVI involves rating each item based on the relevancy scores, in which "1" = not relevant, "2" = somewhat relevant, "3" = quite relevant, and "4" = highly relevant (Lynn, 1986). These relevancy scores are then recoded "1" for ratings of "3 "or "4" and "0" for ratings of "1" or "2" (Lynn, 1986). From the scores by the raters, a value of CVI needs to be calculated. One of the methods to calculate the value is based on the average (S-CVI/Ave) (Polit and Beck, 2006). The S-CVI/Ave is more commonly used because of its flexibility in considering the possibility of disagreement by averaging the agreement and disagreement of the tested items in the computation (Polit and Beck, 2006; Polit et al., 2007). Davis (1992) recommended 0.8 and above as the acceptable value for the S-CVI/Ave. For the item with a value lower than 0.8, the item needs a revision, usually form the consensus of the raters (Almanasreh et al., 2019; Polit et al., 2007).

2.1.1(b) Construct validity

Construct validity is the evidence to a group of variables represents the construct to be measured (Souza et al., 2017). Apart from structural or factorial analysis, the other common approach to determine the construct validity is by using the hypothesis testing strategy (Souza et al., 2017), in which the known-group method is usually used (Souza et al., 2017; Waltz et al., 2010). Through this method, the scores of two distinct groups, usually the normal group versus the disordered group, will be compared to confirm the construct validity (Cronbach and Meehl, 1955; Devon et al.,

2007; Souza et al., 2017). Based on the theoretical foundation, this method provides the indication of the score's direction (Devon et al., 2007), in which if an increase in the scores suggests an increase in the severity of the disorder, the disordered group should record higher scores than the normal group, vice versa. The requirement for the construct validity of the tool is considered achieved if the scores of the tool can discriminate between the normal and disordered groups (Devon et al., 2007; Souza et al., 2017).

2.1.1(c) Concurrent validity

Concurrent validity is the evidence that shows a positive relationship between a translated tool and an established tool (Aaronson et al., 2002; Souza et al., 2017). The established tool must measure the same target (Guillemin, 1995; Souza et al., 2017), and it is usually chosen based on several criteria such as the convenience of use, administration time, and cost (Kimberlin and Winterstein, 2008; Souza et al., 2017). The concurrent validity analysis is performed by comparing the measurement outcomes of the translated tool with those of the established tool, which are obtained from the same group of a sample at the same time (Polit, 2015; Souza et al., 2017). For it to be considered concurrently validated, the measurement outcomes of the translated tool and established tool must show an acceptable positive correlation (Devon et al., 2007; Souza et al., 2017).

2.1.2 Reliability

Reliability is the ability to reproduce a consistent, accurate, and stable result even though it was carried out at different times, places, or by different individuals (Aaronson et al., 2002; Carding et al., 2009; Souza et al., 2017) and free from random error or unwanted variation (Aaronson et al., 2002; Souza et al., 2017). There are several types of reliability, such as internal consistency, intra-rater reliability, and

intra-rater reliability (Aaronson et al., 2002; Devon et al., 2007; Guillemin, 1995). The common types of reliability analyses are inter-rater reliability and intra-rater reliability (Aaronson et al., 2002; Polit, 2015; Souza et al., 2017).

2.1.2(a) Inter-rater reliability

Inter-rater reliability is the consistency of different raters in rating the behaviors by using the same tool (Souza et al., 2017; Waltz et al., 2010). A minimum of two independent raters are needed in this analysis (Polit, 2015; Souza et al., 2017; Waltz et al., 2010). Each of the independent raters performs the rating by using the same tool on the same group of a sample. The agreement of rating scores between the raters is analyzed to determine the inter-rater reliability of the tool (Waltz et al., 2010). The tool is considered to have appropriate inter-rater reliability if the rating scores of the different raters achieved an acceptable level.

2.1.2(b) Intra-rater reliability

Intra-rater reliability refers to the consistency of the raters in assigning the scores to the same sample at two different times (Kimberlin and Winterstein, 2008; Souza et al., 2017; Waltz et al., 2010). The interval between the ratings at the two different times is crucial to ensure sufficient duration to eliminate the memory (i.e., the responses from the first rating session) (Devon et al., 2007; Kimberlin and Winterstein, 2008; Polit, 2014; Trochim and Donnelly, 2001). Commonly, one week (Polit, 2014) to two weeks (Devon et al., 2007; Polit, 2014; Souza et al., 2017; Waltz et al., 2010) are used as the time interval between the first rating session and the second rating session. The intra-rater reliability was analyzed by examining the agreement between the two rating scores on the same sample (Devon et al., 2007; Souza et al., 2017). The rating scores by the same rater at different rating sessions that achieved an acceptable level suggest that the tool has adequate intra-rater reliability.

2.2 Voice Disorder

According to the American Speech-Language-Hearing Association (ASHA) (n.d.), an individual is considered to have a voice disorder when the vocal quality, pitch, or loudness of the individual is inappropriate for the individual's age, gender, culture, or geographic region.

Based on ASHA (n.d.), voice disorders are classified into:

i. organic voice disorder that results from an alteration in the respiratory, laryngeal, or vocal tract mechanism, which can be contributed to:

a. structural changes in vocal mechanisms, such as vocal fold nodules and vocal fold edema.

b. neurogenic problems involving issues with laryngeal innervation that affects the functioning of the vocal mechanism, such as vocal fold paralysis and spasmodic dysphonia.

ii. functional voice disorder, which refers to issues caused by the inappropriate or inefficient use of vocal mechanisms when the physical structure of vocal folds is normal, such as vocal fatigue and muscle tension dysphonia.

iii. psychogenic voice disorder, which is caused by psychological stress such as psychogenic voice disorder or conversion aphonia.

2.3 Auditory-Perceptual Rating of Voice

A voice assessment provides information that helps the clinician to manage voice disorder by determining the appropriate diagnosis, plan, and direction of therapy goals, which would improve the quality of life of a patient (Andrews, 2006; Patel et al., 2018). Due to the complexity of voice production, which involves respiratory, phonatory, and resonatory mechanisms, no single assessment method can completely

measure voice (Roy et al., 2013). In clinical practice, voice assessment includes a combination of procedures, namely evaluation of structure and function of the larynx/vocal folds, auditory-perceptual rating of voice, analysis of voice acoustic, evaluation of breathing for phonation, and self-perception of vocal health (Aronson and Bless, 2009; Colton et al., 2011; Roy et al., 2013; Sataloff, 2017).

Auditory-perceptual rating of voice is an evaluation procedure, in which the clinician listens to the voice samples to judge the vocal parameters based on measurement scales (Boone et al., 2014; Carding et al., 2009; Zraick et al., 2011). In general, the auditory-perceptual rating of voice consists of the following components:

i. Voice samples

The voice samples usually consist of a single sound, sentence, or running speech, in any combination. Several vocal tasks were used to collect the voice samples, such as prolongation of vowels to obtain single sounds (e.g., Brinca et al., 2015; Gerratt et al., 2016; Kempster et al., 2009), imitation or reading to elicit sentences (e.g., Gerratt et al., 2016; Kempster et al., 2009), and reading passage (e.g., Bhuta et al., 2004; Brinca et al., 2015; Zraick, Wendel, et al., 2005), describing a picture (e.g., Titze, 1995; Zraick, Wendel, et al., 2005) or answering a question (e.g., Brinca et al., 2015; Kempster et al., 2009) to gather running speech.

ii. Vocal parameters

Varied types of vocal parameters have been used to characterize voice. The main vocal parameters are voice quality. As described by Kempster et al. (2009) and Hirano (1981), voice quality commonly includes:

a. Global severity: The global, integrated impression of voice deviance,

b. Roughness: The perceived irregularity in the voicing source,

c. Breathiness: The audible air escape in the voice,

d. Strain: The perception of excessive vocal effort (hyperfunction),

e. Asthenia: Weakness or lack of power.

Apart from the above voice qualities, other vocal parameters are also included in an auditory-perceptual rating of voice, such as vocal pitch, i.e., the perceptual correlate of the fundamental frequency, and vocal loudness, i.e., the perceptual correlate of the sound intensity (Kempster et al., 2009).

iii. Measurement scales

Rating of the vocal parameters is performed on measurement scales, which are usually either in the form of a visual analog scale (VAS) or an equal-appearing interval scale (EAIS) (Kreiman and Gerratt, 1998; Ma and Yiu, 2007; Yiu and Ng, 2004). VAS requires the rater to assign a number that corresponds to the magnitude of the rating scale (Ma and Yiu, 2007). EAIS, on the other hand, requires the raters to assign a numerical or descriptor value that is arranged in intervals along a linear continuum (Ma and Yiu, 2007). Both scales are valid and reliable in evaluating the voice (Ma and Yiu, 2007). However, because it does not require a manual measurement to identify the scores, EAIS is considered to be more practical and less time-consuming than VAS (Ma and Yiu, 2007).

2.4 Tools for Auditory-Perceptual Rating of Voice

Several tools for the auditory-perceptual rating of voice have been introduced, such as the GRBAS Scale (Hirano, 1981), Vocal Profile Analysis Scheme (Laver et al., 1981), Buffalo Voice Profile (Wilson, 1987), and CAPE-V (Kempster et al., 2009). Although each tool has its merit, the GRBAS Scale and CAPE-V were the most commonly utilized tools in clinical practice (Boone et al., 2014; Mcalister et al., 2020).

2.4.1 The GRBAS Scale

The GRBAS Scale was developed by the Japanese Society of Logopedics and Phoniatrics (Hirano, 1981). As it provides meaningful vocal quality information (Carding et al., 2000), the GRBAS Scale is one of the most commonly used auditory-perceptual rating of voice tools in clinical practice (Mcalister et al., 2020). In the procedures, the voice samples are rated according to the five main voice qualities, i.e., overall severity, roughness, breathiness, asthenia, and strain, which are represented by G, R, B, A, and S, respectively. The degree of deviancy for each voice quality is judged based on an EAIS (i.e., a four-point Likert scale), where "0" indicates no deviancy, "1" implies slight/mild deviancy, "2" represents moderate deviancy, and "3" suggesting severe/extreme deviancy (Hirano, 1981). The result of the GRBAS Scale is presented individually according to the rated point for each of the voice qualities.

Albeit widely used, the GRBAS Scale is lacking in standardized administration procedures and interpretation guidelines (Kempster et al., 2009). The lack of standardized protocol resulted in inconsistent types of voice samples used, such as single sounds (i.e., Yamaguchi et al., 2003), running speech (i.e., Tedla et al., 2016; Tezcaner et al., 2009; Vilaseca et al., 2008), or a combination of single sounds and running speech (i.e., Brinca et al., 2015; De Bodt et al., 1997; Wuyts et al., 1999). Even though the four-point Likert scale (i.e., "0", "1", "2", and "3") is used in presenting the results, the interpretations of the scale are found to be varied. For example, "0" may represent no abnormality (i.e., Yamaguchi et al., 2003), normal (i.e., Brinca et al., 2015; Tezcaner et al., 2009; Vilaseca et al., 2008), non-hoarse/absent of hoarseness (i.e., De Bodt et al., 1997; Jesus et al., 2017), or neutral (i.e., Dos Santos et al., 2019). Meanwhile, scales "1" until "3", respectively, may be interpreted as slight, moderate, and severe (i.e., De Bodt et al., 1997; Jesus et al., 2017; Wuyts et al., 1999), slight

voice problem, moderate, and severe (i.e., Vilaseca et al., 2008), slight disturbance, moderate disturbance, and severe disturbance (i.e., Tezcaner et al., 2009).

2.4.2 The Consensus Auditory-Perceptual Evaluation of Voice

With specific vocal tasks, procedures, and ratings, the CAPE-V (Appendix B) was developed to promote a standardized tool for an auditory-perceptual rating of voice (Kempster et al., 2009). In the CAPE-V, three types of voice samples are included single sounds (i.e., /a/ and /i/), six sentences, each following different phonetic contexts (Table 2.1), and 20-seconds of running speech answering a question "Tell me about your voice problem" or "Tell me how your voice is functioning".

Table 2.1 The sentence for each phonetic context in the CAPE-V

	Sentences	Phonetic Context
(a)	The blue spot is on the key again	Production of all vowels in the particular language
(b)	How hard did he hit him?	Elicitation of easy onset
(c)	We were away a year ago	Production of voiced sounds
(d)	We eat eggs every Easter	Elicitation of hard glottal attack
(e)	My mama makes lemon jam	Production of nasal sounds in the particular language
(f)	Peter will keep at the peak	Production of voiceless plosive sounds in the particular language

By listening to all three types of voice samples, the voice is judged based on the vocal parameters of main voice qualities (i.e., overall severity, roughness, breathiness, and strain), pitch, and loudness (Kempster, 2007; Kempster et al., 2009). If the need arises, the CAPE-V also provides a rating of optional voice qualities (e.g., aphonia, falsetto, and spasm) and resonance (Kempster, 2007; Kempster et al., 2009).

The vocal parameters are measured using a 100-millimeters VAS, which corresponds to an increased deviancy of voice. The perception of deviance of each vocal parameter is indicated by a strike along the 100-millimeter VAS. The length of the line, where the strike is located, is then measured by a ruler. The length of the line

in millimeters represents the score for the deviancy of voice. The longer the length, i.e., the higher the score, the more severe the perceived vocal parameter. The result of the CAPE-V is presented individually according to the measured length, i.e., the score for each of the voice qualities.

2.5 Cross-cultural adaptation of CAPE-V

Given that it provides standard administration and documentation procedures, as well as a valid and reliable tool, the CAPE-V has been adapted into many languages, such as Brazilian Portuguese (i.e., Behlau et al., 2020; Nemr et al., 2012, 2016), European Portuguese (i.e., De Almeida et al., 2016; Jesus, Barney, Couto, et al., 2009; Jesus et al., 2017), Hindi (i.e., Isha Baheti, 2019; Joshi et al., 2020), Italian (i.e., Mozzanica et al., 2014), Kannada (i.e., Gunjawate et al., 2020), Mandarin (i.e., Chen et al., 2018), Spanish (i.e., Núñez-Batalla et al., 2015), and Turkish (i.e., Ertan-Schlüter et al., 2020; Özcebe et al., 2017).

Apart from the recommended processes, i.e., translation into intended language, examination of validity (i.e., content validity, construct validity, and concurrent validity), and examination of reliability (i.e., inter-rater reliability and intra-rater reliability), it was found that several studies (i.e., Chen et al., 2018; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Isha Baheti, 2019) also employed some modifications of the original CAPE-V.

2.5.1 Modification

Two types of modifications were employed by the previous studies, i.e., modifications of the rating scale and question in the free conversation task. Chen et al. (2018) modified the scale from the VAS to EAIS, in which the 100-millimeter VAS was transformed into a 101-point EAIS. This modification allowed the clinician to

easily read the score without the need to measure the line manually using a ruler. Meanwhile, a few studies (i.e., Chen et al., 2018; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Isha Baheti, 2019) modified the questions in the free conversation task from the topic of voice problem to more general topics, such as "Tell me what you did after getting up", "Tell me about the place where you grew up", "Tell me about your daily routine", and "Tell me about your family", to elicit more neutral and longer utterances (Chen et al., 2018; Zraick et al., 2011).

2.5.2 Translation

A review of the published literature revealed that not all the adaptation studies presented information on the translation process. Even among those that did, the information was rather limited. These studies only reported information on the translation of the six sentences in the sentence production task (i.e., Chen et al., 2018; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Gunjawate et al., 2020; Isha Baheti, 2019; Joshi et al., 2020; Núñez-Batalla et al., 2015; Özcebe et al., 2017) and the question in the free conversation task (i.e., Chen et al., 2018; De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Isha Baheti, 2019). None of the studies completely reported information based on the recommended translation process, which should involve forward translation and backward translation.

2.5.3 Validity

2.5.3(a) Content validity

Similar to that on translation, limited information was also evident in the previous literature on content validity. The only available information was confined to the inspection of the six sentences in the sentence production task according to the phonetic contexts by linguists (i.e., De Almeida et al., 2016; Ertan-Schlüter et al., 2020; Joshi et al., 2020), phoniatricians (i.e., Chen et al., 2018), and SLTs (i.e., Ertan-

Schlüter et al., 2020; Núñez-Batalla et al., 2015). Again, none of the previous studies performed the complete recommended procedure to examine the content validity that should involve the analysis of the domain of the items by using an appropriate index calculation.

2.5.3(b) Construct validity

All the previous studies, which analyzed the construct validity, used the known-group method, comparing the normal and disordered voice groups. In all these studies, only six vocal parameters, i.e., overall severity, roughness, breathiness, strain, pitch, and loudness, were involved in the analysis. Two types of measures were used to compare the vocal parameters between the normal and disordered voice groups, i.e., the individual score from each rater and the average score from all raters. As shown in Table 2.2, there were significant differences in almost all the analyzed vocal parameters between the normal and disordered voice groups, regardless of the types of measures.

Table 2.2 The results of the construct validity from previous studies.

Study	Type of Measure	Vocal Parameter					
		Overall Severity	**Roughness**	**Breathiness**	**Strain**	**Pitch**	**Loudness**
Mozzanica et al. (2014)	Individual score	√	√	√	√	√	√
De Almeida et al. (2016)	Average score	√	√	√	X	√	√
Nemr (2016)	Average score	√	√	√	√	√	√
Özcebe et al. (2017)	Individual score	√	√	√	√	√	√
Gunjawate et al. (2020)	Average score	√	√	√	√	√	X
Ertan-Schlüter et al. (2020)	Individual score	√	√	√	√	√	√

√ = significant, $p<0.05$
X = not significant

2.5.3(c) Concurrent validity

All the previous studies examined the concurrent validity by inspecting the relationship between the translated CAPE-V and the GRBAS Scale. Meanwhile, almost all these studies involved the four corresponding vocal parameters of the two tools, i.e., overall severity/grade, roughness, breathiness, and strain (i.e., Behlau et al., 2020; De Almeida et al., 2019; Ertan-Schlüter et al., 2020; Gunjawate et al., 2020; Mozzanica et al., 2014; Núñez-Batalla et al., 2015; Özcebe et al., 2017; Zraick et al., 2011). Except for Joshi et al. (2020), the concurrent validity analyses of all previous studies involved the combination of normal and disordered voice participants. Apart from the average score, the average correlation value, i.e., from the correlation score of all raters, was also used to determine the concurrent validity. In summary, the results on the concurrent validity of the previous studies indicated positive correlations in all vocal parameters between the translated CAPE-V and GRBAS Scale.

Table 2.3 The results of the concurrent validity from previous studies.

Study	Statistical analysis	Type of Measure	Vocal Parameter			
			Overall Severity	**Roughness**	**Breathiness**	**Strain**
Zraick et al. (2011)	Multi-serial correlation	Average correlation value	0.80	0.76	0.78	0.77
De Almeida et al. (2016)	Multi-serial correlation	Average score	0.95	0.89	0.90	0.47
Mozzanica et al. (2014)	Spearman correlation	Average correlation value	0.92	0.84	0.87	0.79
Özcebe et al. (2017)	Spearman correlation	Average correlation value	0.80	0.62	0.67	0.67
Gunjawate et al. (2020)	Spearman correlation	Not available	0.96	0.94	0.95	0.89
Ertan-Schlüter et al. (2020)	Spearman correlation	Average correlation value	0.85	0.82	0.77	0.66
Behlau et al. (2020)	Spearman correlation	Average score	0.95	0.94	0.92	0.83
Núñez-Batalla et al. (2015)	Intraclass correlation coefficient	Average correlation value	0.87	0.85	0.61	0.84
Joshi et al. (2020)	Pearson correlation	Not available				
Normal voice group			0.75	0.76	0.63	NA
Disordered voice group			0.96	0.84	0.87	0.79

NA = Not Applicable

2.5.4 Reliability

2.5.4(a) Inter-rater reliability

The previous studies used a minimum of two independent raters in the analysis of inter-rater reliability. In all the previous studies, six vocal parameters of the CAPE-V, i.e., overall severity, roughness, breathiness, strain, pitch, and loudness, were involved in the analyses. As summarized in Table 2.4, all the studies used the intraclass correlation coefficient (ICC) in the analyses, with the majority of them scoring a value above 0.8.

Table 2.4 The results of the inter-rater reliability from previous studies.

Study	Number of Raters	Vocal Parameter					
		Overall severity	**Roughness**	**Breathiness**	**Strain**	**Pitch**	**Loudness**
Zraick et al. (2011)	21	0.76	0.62	0.60	0.56	0.28	0.54
Mozzanica et al. (2014)	3	0.92	0.91	0.90	0.76	0.83	0.82
De Almeida et al. (2016)	14	0.96	0.92	0.95	0.84	0.86	0.90
Özcebe et al. (2017)	4	0.90	0.81	0.84	0.80	0.88	0.81
Chen et al. (2018)	4	0.81	0.80	0.76	0.59	0.57	0.70
Gunjawate et al. (2020)	3	0.97	0.97	0.98	0.97	0.98	0.96
Ertan-Schlüter et al. (2020)	2	0.95	0.91	0.93	0.88	0.89	0.92
Behlau et al. (2020)	9	0.86	0.67	0.71	0.32	0.26	0.68