Validity and reliability of the 2nd European Portuguese version of the "*Consensus Auditory-Perceptual Evaluation of Voice*" (II EP CAPE-V)

By

SANCHA C. DE ALMEIDA

Master thesis

A thesis submitted in partial fulfillment of the requirement for the degree of Master in Science at the Health Science School of Polytechnic Institute of Setúbal

May 2016

IPS Instituto Politécnico de Setúbal
Escola Superior de Saúde

Validity and reliability of the 2nd European Portuguese version of the "*Consensus Auditory-Perceptual Evaluation of Voice*" (II EP CAPE-V)

By

SANCHA C. DE ALMEIDA, nº 130522006

Unit course: "Project work II"

Voice Disorders and (Re)Habilitation Master Program

Health Science School of Polytechnic Institute of Setúbal – Portugal

Supervisor: **Ana P. Mendes,** Ph.D., CCC-SLP

Co-supervisor: **Gail B. Kempster,** Ph.D., CCC-SLP

Setúbal, May 25th 2016

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor and co-advisor who gave their time, guidance, encouragement, and patience to make this thesis possible. Professor Ana P. Mendes was who first encouraged me to go further; she inspired me to continuously ask questions and searching for valid and reliable work in the field of voice. Thank very much with all my heart for giving me your unconditional support! Professor Gail B. Kempster accepted to become my co-advisor, without knowing me personally; she gave me her support and expertise in the area of auditory-perceptual assessment of voice. Thank very much for having trusted in me and in my abilities.

I would like to thank Professor Fernando Martins who passionately explained European Portuguese phonetics and helped me with the content validity review. Thanks to my colleagues Ana Paula Almeida, Rosa Henriques, Mariana Pinheiro, and Elisabete Afonso who started to work on CAPE-V translation few years ago. Thanks also to Lisa Carvalho e Silva, Mónica Carvalho e Silva and Carlos Ibrahim, who helped me with the translation of the CAPE-V sentences.

I'm extremely grateful to the following Speech and Language Pathologists, who freely took the time to attend the voice rating sessions: Aira Rodrigues, David Guerreiro, Inês Moura, Joana Assunção, João Frataria, Leonor Fontes, Luísa Pacheco Nobre, Mafalda Almeida, Maria Filomena Gonçalves, Mariana Moldão, Miriam Moreira, Sónia Lima, Soraia Ibrahim, Tânia Constantino, and Teresa Rosado. Without them, this study would not have been possible.

A special thank you to Professor Margarida Lemos and Professor Maria Fátima Salgueiro for their precious time and availability to explain and assist with statistics.

Thanks to the ENT team, individuals whom I'm lucky to work with, especially to Dr. António Larroudé, Dr. Sara Viana Baptista, and Dr. Rita Ferreira who were always present during this time, sharing their knowledge and supporting me to fly high.

A special thanks to Soraia Ibrahim and her husband, who patiently gave me their time and valuable insights, no matter what time it was.

Lastly, I would also like to thank my family and friends for their immeasurable understanding and support. Thank you all for believing in me.

ABSTRACT

Introduction: Auditory-perceptual evaluation of voice is a part of a multidimensional voice evaluation, and is claimed to be "*golden standard*". The "*Consensus Auditory-Perceptual Evaluation of Voice*" (CAPE-V) has been demonstrated to be a valid and reliable instrument for voice evaluation, when applied in both clinical and scientific research fields. The CAPE-V was first translated into European Portuguese (EP) (Jesus et al., 2009) however it revealed some validity and reliability problems. The purpose of this study was to assure a valid and reliable EP version of CAPE-V. This resulted in the 2nd EP version of CAPE-V (II EP CAPE-V), with permission granted by ASHA.

Method: This was a transversal, observational, descriptive, and comparative study. 14 Speech-language pathologists (SLPs) voice experts (>5 years of clinical practice), rated a total of 26 voice samples produced by 10 males (mean age=45) and 10 females (mean age=43) classified into two groups: a control group (n=10) and a dysphonic group (n=10), with subjects matched for age and gender. All voice samples were rated in one session with the II EP CAPE-V, and in a second session one week later with GRBAS. Content validity was supported by 6 new sentences conceptualized and adapted to EP linguistic and cultural context according to the rationale outlined in the original CAPE-V protocol. For construct validity analysis, an independent samples *t*-test (α=.05) was performed for all vocal parameter. Concurrent validity was estimated with the multi-serial correlation coefficient between II EP CAPE-V and GRBAS parameters (r>.70). Reliability was performed for all vocal parameters. Inter-rater reliability was determined by ICC, and intra-rater reliability by Pearson's correlation coefficient (r>.70).

Results/conclusion: Content validity was assured by an EP linguistic expert, who reviewed the six new sentences. Construct validity was obtained for all voical parameters (p<.05), except for strain (p=.52). Concurrent validity had high correlations (r>.89) for overall severity/grade, roughness, and breathiness parameters. High inter-rater reliability (ICC>.84) was obtained for all parameters. Intra-rater reliability was high (r>.87) for overall severity, breathiness, and pitch; good (r=.73) for strain; and moderate (r>.69) for roughness and loudness parameters. The II EP CAPE-V is a valid and reliable instrument for auditory-perceptual evaluation, with all psychometric characteristics established.

Key words: CAPE-V, voice evaluation, auditory-perceptual evaluation, dysphonia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AE	American English
ASHA	American Speech and Hearing Association
BP	Brazilian Portuguese
CAPE-V	Consensus Auditory-Perceptual Evaluation of Voice
CG	Control group
DG	Dysphonic group
EAI	Equal appearing interval
ENT	Ear nose and throat
EP	European Portuguese
GRBAS	Grade; rough; breathy; asthenic; strained
ICC	Intraclass correlation coefficient
II EP CAPE-V	2nd European Portuguese version of CAPE-V
IT	Italian
LNO	Limitations were not observed
NA	Not available
RCSLT	Royal College of Speech and Language Therapist
SACMOT	Scientific Advisory Committee of the Medical Outcomes Trust
SLP	Speech-language pathologist
SP	Spanish
VAS	Visual analog scale
VF	Vocal Fold
VQ	Voice quality

I. INTRODUCTION

The present master thesis was developed in the course unit "*Project work II*" of the Voice Disorders and (Re)Habilitation Master's program at Health Science School of Polytechnic Institute of Setúbal – Portugal. This study is named "Validity and reliability of the 2nd European Portuguese version of the *"Consensus Auditory-Perceptual Evaluation of Voice*" (II EP CAPE-V)". It was supervised by Professor Ana P. Mendes and co-supervised by Professor Gail B. Kempster.

Auditory-perception plays an invaluable role in voice field. Usually patients seek treatment because their voices sound perceptually different than the normal. They decide if the treatment has been successful based upon their voices 'sounds better than before' (Awan & Lawson, 2009; Kreiman, Gerratt, Kempster, Erman, Berke, 1993; Shewell, 1998). Voice intervention outcomes (e.g. surgical, therapy) are measured by different types of voice evaluation methods such as: auditory-perceptual, acoustic, aerodynamic, laryngoscopic, and self-evaluation (Barsties & De Bodt, 2015; Kelchner et al., 2010; Mehta & Hillman, 2008; MgGlashan & Fourcin, 2008; Speyer, 2008).

Auditory-perceptual voice evaluation is claimed to be the *golden standard* for voice evaluation (Oates, 2009; Speyer, 2008). This evaluation is based on the listener's perception of the different vocal parameters or quality aspects present in normal or dysphonic voice samples (Carding, Carlson, Epstein, Mathieson & Shewell, 2000; Carding, Wilson, MacKenzie & Deary, 2009). The auditory-perceptual analysis of voice quality is often considered to be subjective and influenced by several factors related to listeners, voice stimuli and the rating scale applied (Bassich & Ludlow, 1986; Bele, 2005; Brinca, Batista, Tavares, Pinto & Araújo, 2015; Eadie & Baylor, 2006; Eadie, Boven, Stubbs & Giannini, 2010; Kreiman & Gerratt, 1998; Kreiman, Gerratt & Precoda, 1990; Kreiman et al., 1993; Kreiman, Gerratt, Precoda & Berke, 1992; Maryn & Roy, 2012; Oates, 2009; Sofranko & Prosek, 2012; Wuyts, De Bodt & Van de Heyning, 1999; Zraick, Wendel & Smith-Olinde, 2005).

According to the current standards of evidence-based medicine, any instrument of health status evaluation should demonstrate having evidence of validity and reliability in order to be clinically useful (Aaronson et al., 2002; Carding et al. 2009). Different

instruments were designed to promote a standardized auditory-perceptual voice evaluation. GRBAS (Hirano, 1981) and "*Consensus Auditory-Perceptual Evaluation of Voice*" (CAPE-V) (ASHA, 2006) are two widely used instruments in clinical and research fields; their validity and reliability are well reported. The validity and reliability of these auditory-perceptual instruments had already been studied when they were first translated to European Portuguese (EP) (Freitas, Pestana, Almeida & Ferreira., 2014; Jesus, Barney, Sá Couto, Vilarinho & Correia, 2009b; Jesus, Barney, Santos, Caetano, Jorge & Sá Couto, 2009a). However, the first EP version of CAPE-V revealed some validity and reliability problems (Jesus et al., 2009b; Jesus et al., 2009a). In order for the Portuguese voice clinicians to be able to evaluate and treat voice patients, as well as to compare and share results from Portuguese research or clinical practice with other national or international research and clinical colleagues, a new translation of CAPE-V into EP was needed. The purpose of the present study was to develop a valid and reliable EP version of the AE (American English) 2nd edition of CAPE-V (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer & Hillman, 2009).

This study reviews literature that explains auditory-perceptual voice evaluation according to the principles for such instruments – validity and reliability. It also names different instruments available for auditory-perceptual evaluation; of the instruments mentioned, the GRBAS and CAPE-V are explained in detail. This thesis presents the validity and reliability of auditory-perceptual evaluation, and the CAPE-V validity and reliability results for the second EP translation. The next chapter explains the methodology used in this study, and it is followed by the results obtained and by a discussion with limitations present. This paper finishes with the study's conclusions and comments regarding future research.

II. REVIEW OF THE LITERATURE AND RESEARCH QUESTIONS

2.1. Auditory-Perceptual Voice Evaluation

"*Every human voice is unique because of anatomical, physiological, psychological, cultural, sociolinguistic and behavioral factors*" (MgGlashan & Fourcin, 2008, pp. 2171). Voice quality (VQ) is a perceptual phenomenon (Barsties & De Bodt, 2015; Patel & Shrivastav, 2007) that can be translated as the listener's subconscious reaction to a voice's acoustic signal (Brinca et al., 2015). It can be understood as audible sound resultant from different factors, and it can be described using terms such as breathiness, roughness, or harshness (Guimarães, 2007; Speyer, 2008).

Up to the present time, existing literature has not agreed upon a definition of the term "normal voice"; however this term can be related to ordinary speaking voices that are not dysfunctional (Bele, 2007). The term dysphonia is used when a VQ disorder exists, manifesting itself as a disturbance in vocal emission that results in natural voice production (Behlau & Pontes, 1995). According to Otolaryngology Head and Neck Surgery clinical practice guidelines (Schwartz et al., 2009, pp.S2) this term "*is defined as a disorder characterized by altered vocal quality, pitch, loudness, or vocal effort that impairs communication or reduces voice-related quality of life*". Usually, a voice is labeled as disordered when one or more perceptual features of VQ are audibly dissimilar to those of people of the same sex, age, and culture (MgGlashan & Fourcin, 2008). This term should also be used when any deviation in VQ is perceived, whether it concerns pitch, loudness, timbre, or rhythmic and prosodic features (Dejonckere et al., 2001).

A voice disorder can be a result of structural, inflammatory, traumatic, systemic, non-laryngeal aerogestive, psychiatric, psychological, neurologic, neuromuscular, or from any other disorders that may affect the voice production system (Carding & Mathieson, 2008; Royal College of Speech & Language Therapists [RCSLT], 2009; Verdolini, Rosen & Branski, 2006). The diagnosis of a voice disorder should involve a series of specific procedures that include clinical diagnosis and VQ assessment, which can only be performed by qualified professionals (Ghirardi, Ferreira, Giannini & Latorre, 2013).

Measuring the VQ is important in the clinical evaluation and rehabilitation of patients with dysphonic voices (Patel & Shrivastav, 2007). Hitherto, there is no instrument nor value that can quantify or characterize a human voice disorder by itself (Kelchner et al., 2010; Shewell, 1998). Voice evaluation is still a multifactorial process, where different aspects of voice production are assessed through the auditory-perceptual evaluation of VQ, acoustic evaluation of voice sound production, aerodynamic evaluation of subglottal air pressures and glottal airflow during voicing, endoscopic imaging of vocal fold (VF) tissue vibration, quality-of-life measurements, and self-evaluation by the patient (Barsties & De Bodt, 2015; Kelchner et al., 2010; Mehta & Hillman, 2008; MgGlashan & Fourcin, 2008; Speyer, 2008).

Auditory-perceptual evaluation of voice is part of multidimensional voice evaluation (Carding et al., 2009) and is one of the most traditional approaches in VQ analysis (Nemr et al., 2012). It is considered as the "*golden standard*" for documenting voice disorders (Speyer, 2008; Oates, 2009). This type of evaluation is non-invasive, thus comfortable to the patient; it is succinct, quick to perform, and low cost. The results can be easily communicated between clinicians. All of these factors makes it a valued procedure used worldwide (Carding et al., 2000; Carding et al., 2009; Oates, 2009; Sáenz-Lechón, Godino-Llorente, Osma-Ruiz, Blanco-Velasco, Cruz-Roldán, 2006; Wuyts, De Bodt & Van de Heyning 1999).

Perception is a mental construction resulting from processing of the available present information added to our past internal standards (Ghio, Révis, Merienne & Giovanni, 2013). Any auditory stimulus is an interaction between an acoustic voice stimulus and a listener's response to that stimulus. VQ is the perceptual response to an acoustic voice signal (Kreiman & Gerratt, 1998). The auditory-perceptual evaluation assesses VQ based on the auditory impression that a listener has when listening to a disordered or normal voice (Nemr et al., 2012). This process involves an expert listener judging a voice sample across various vocal parameters (Carding et al., 2000; Carding et al., 2009), assessing and grading their severity on a predetermined scale (Bless et al., 1992).

Usually this type of voice evaluation is conducted to provide clinical information about the type and severity of the dysphonia (Carding et al., 2000; Ghio et al. 2013). It allows the clinician to establish a baseline and to measure an individual's progress

throughout intervention (Oates, 2009; RCSLT, 2009). VQ assessment is relevant for studies of surgical treatment outcomes and behavioral approaches to management of voice disorders (Gould, Waugh, Carding & Drinnan, 2012; Karnell, Melton, Childes, Coleman, Dailey & Hoffman, 2007). The auditory-perceptual results added to the patient's complaints, history of dysphonia, and vocal self-assessment enables the speech-language pathologist (SLP) to plan a series of activities to improve both the VQ and the quality of life of the individuals suffering from voice disorders (Behrman, 2005; Carding et al., 2000).

Auditory-perceptual voice evaluation is particularly relevant when assessing patients with severe dysphonia. In these cases the voice signal is highly aperiodic which limits acoustic voice analysis (Kelchner et al., 2010). Despite all of the advances in acoustic voice analysis, the accuracy of acoustic measures is limited as a result of the difficulty of accurate determination of the fundamental frequency (f_o) (Leong, Hawkshaw, Dentchev, Gupta, Lurie & Sataloff, 2013; Mehta & Hillman, 2008).

However, auditory-perceptual VQ evaluation is a difficult task (Bassich & Ludlow, 1986) because it is subjective and it can be influenced by different factors such as: listeners' internal standards, listeners' background experience, listeners' training, type of rating scale, and type of voice sample (Awan & Lawson, 2009; Bassich & Ludlow, 1986; Eadie et al., 2010; Iwarsson & Petersen, 2012; Kreiman et al., 1990; Kreiman et al., 1993; Kreiman et al., 1992; Kreiman, Vanlancker-Sidtis & Gerratt, 2004; Law et al., 2012; Oates, 2009; Shrivastav, Sapienza & Nandur, 2005; Sofranko & Prosek, 2012).

Auditory-perceptual evaluation relies on comparing one voice with another or comparing different voice productions produced by the same subject (Bele, 2005; Fex, 1992). These tasks can lead to poor sensitivity and poor agreement across individual raters (Gerratt, Kreiman, Antonanzas-Barroso & Berke, 1993), limiting the validity and reliability of the auditory-perceptual results.

2.1.1. Perceptual Rating Scales

Rating VQ is mainly a bottom-up perception process in which listeners categorize voices based on a voice sample heard, interpreting acoustic cues detected perceptually (Ghio et al., 2013). When VQ is measured through rating scales on particular aspects of

quality, it is assumed that the overall impression of a voice received by a listener could be decomposed into several perceptually distinct aspects corresponding to various terms, such as breathiness and roughness (Kreiman & Gerratt, 1998).

Accurate auditory-perceptual judgments of VQ can be made if the correct tools are available (Gould et al., 2012). Measurement tools should remain constant across listeners and voice samples, so that different listeners can use the scales in the same way, and the measurements of different voices can be meaningfully compared. This way voice quality features can be treated as attributes of the voice signal itself, rather than as the product of a listener's perception (Kreiman & Gerratt, 1998).

For clinical and research purposes, a voice outcome measurement tool should be valid, reliable, and sensitive to change (Carding et al., 2009). In an effort to standardize auditory-perceptual voice evaluation, different schemes and scales specifically designed for this purpose have been developed such as the GRBAS scale (Hirano, 1981), GIRBAS (Dejonckere, Remacle, Fresnel-Elbaz, Woisnard, Crevier-Buchman, 1996), RASAT (Pinho & Pontes, 2002), RASATI (Pinho & Pontes, 2008), GRBASH (Nemr & Lehn, 2010), (I)INFVo (Moerman et al., 2006a, 2006b), Stockholm Voice Evaluation Approach (Hammarberg, 2000), Vocal Profile Analysis Scheme (Laver, Wirz, MacKenzie & Hiller, 1981), Buffalo Voice Profile (Wilson, 1987), and Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) (American Speech-Language-Hearing Association [ASHA], 2006). All of these scales have similarities and differences. They have similarities in the vocal parameters to be judged and their definitions. They differ in the procedures, phonatory tasks, and rating scales in which the auditory-perceptual parameters are judged (summarized in Table 1). However, these factors do not ensure the validity and reliability of these scales and their results (Oates, 2009).

The selection of an auditory-perceptual scale should depend on the clinical and scientific purpose of the evaluation. This requires a careful consideration of the underlying theoretical framework, VQ parameters assessed, and their operational definitions. The type of rating scale, voice sample and recording protocols, and formalized training resources, as well as associated validity and reliability data should be considered for adequate scale or scheme selection (Oates, 2009).

In the next section, two widely used auditory-perceptual evaluation tools the GRBAS and the CAPE-V, will be reviewed in terms of procedures, phonatory tasks, vocal parameters, and rating scales.

Table 1 – Scales and schemes for auditory-perceptual voice evaluations.

Scale/scheme	Authors	Procedures	Phonatory tasks	Auditory-perceptual parameters	Rating scale
GRBAS	Hirano (1981)	Not defined	Not defined	Grade, roughness, breathiness, asthenia, strain	Ordinal scale from 0 to 3
GIRBAS	Dejonckere et al. (1996)	Not defined	Not defined	Grade, instability, roughness, breathiness, asthenia, strain	Ordinal scale from 0 to 3
RASAT	Pinho & Pontes (2002)	Sustaining and speaking aloud, rating procedures, and parameters definition	Sustain /a, i/ and spontaneous speech	Roughness, harshness, breathiness, asthenia, strain	Ordinal scale from 0 to 3, with middle scores of 1.5 and 2.5
RASATI	Pinho & Pontes (2008)	Sustaining and speaking aloud, rating procedures, and parameters definition	Sustain /a, i/ and spontaneous speech	Roughness, harshness, breathiness, asthenia, strain, stability	Ordinal scale from 0 to 3 with middle scores of 1.5 and 2.5
GRBASH	Nemr & Lehn (2010)	Sustaining and speaking aloud, and parameters definition	Sustain /a, i/ and spontaneous speech	Grade, roughness, breathiness, asthenia, strain, harshness	Ordinal scale from 0 to 3 with middle scores 1.5 and 2.5
Impression, Intelligibility, Noise, Fluency, Voicing ((I)INFVo)	Moerman et al. (2006a, 2006b)	Reading aloud, rating procedures, parameters definition, and audio sample example	Utterance of the phonetically rich Dutch text passage	Overall impression, impression of intelligibility, amount of unintended additive noise, fluency, and quality of voicing	VAS divided into 11 cells. The position of the marker can be converted to discrete values from 0–10.
Stockholm Voice Evaluation Approach (SVEA)	Hammarberg (2000)	Not defined	40 seconds of Swedish phonetically balanced text reading	Aphonia/intermittent aphonic, breathy, hyperfunctional/tense, hypofunctional/lax, vocal fry/creacky, rough, gratings/"scrappiness", unstable VQ/pitch, voice breaks, diplophonic, modal/falsetto register, pitch, loudness.	Ordinal scale from 0 to 4
Vocal Profile Analysis Scheme (VPAS)	Laver et al. (1981)	Reading and speaking aloud	Reading and spontaneous speech	31 parameters of VQ, prosodic quality and temporal organization	EAI from 1 to 6

EAI – equal appearing intervals scale; VAS – visual analog scale.

Table 1 (Cont.) – Scales and schemes for auditory-perceptual voice evaluations.

Scale/scheme	Authors	Procedures	Phonatory tasks	Auditory-perceptual parameters	Rating scale
Buffalo III Voice Profile (BVP)	Wilson (1987)	Sustaining, reading and speaking aloud	Sustain vowel not defined, reading, spontaneous speech, and counting.	Laryngeal tone, pitch, loudness, nasal and oral resonance, breath supply, muscles, voice abuse, rate, speech anxiety, speech intelligibility and an overall voice rating.	EAI from 1 to 5
Consensus Auditory Perceptual Evaluation of Voice (CAPE-V)	ASHA (2006)	Sustaining, reading and speaking aloud, procedures for voice recording and rating, and parameters definition	Sustain /a, i/, six sentences, and spontaneous speech.	Overall severity, roughness, breathiness, strain, pitch, loudness	VAS from 0 to 100 mm

EAI – equal appearing intervals scale; VAS – visual analog scale.

2.1.2. GRBAS and CAPE-V

The GRBAS scale was developed by the Japanese Society of Logopedics and Phoniatric to explain the psychoacoustic phenomenon of hoarseness utilizing the Osgood Semantic Differential Technique (Hirano, 1981) (Annex A). This scale is used worldwide in several fields to assess the following VQ aspects:

G – Grade: "degree of abnormality"

R – Rough: "irregularity of fold vibration"

B – Breathy: "air leakage in the glottis"

A – Asthenic: "lack of power"

S – Strained: "hyper functional state".

Each of the vocal parameters are judged using a four point Likert scale from zero (normal) to three (extreme) (Hirano, 1981).

The GRBAS scale is considered as the absolute minimum for voice perceptual evaluation. It has a defined terminology, and it is simple to apply, not offering any discomfort nor inconvenient to the patient or SLP (Carding et al., 2000; Carding et al., 2009). The GRBAS scale is effective for vocal screening and is probably the most compact of all the auditory-perceptual rating systems that can be used easily by all voice team members (De Bodt, Wuyts, Van de Heyning & Croux, 1997; Freitas et al., 2014; Wuyts et al., 1999). However, the GRBAS scale does not provide standardized procedures for evaluation and analysis (Carding et al., 2000; Zraick, Kempster, Connor, Klaben, Bursac & Glaze, 2011). This scale focus on the glottic level, and thus it does not include features such as pitch and loudness nor any other supra-glottic parameter (e.g., resonance) (Carding et al., 2000; Nemr et al., 2012). The four-point ordinal scale used by GRBAS has poor sensitivity for small variations in VQ (Wuyts et al., 1999), and it also cannot be applicable to normal or singing voices (Carding et al., 2000).

The "*Consensus Auditory-Perceptual Evaluation of Voice*" (CAPE-V) (Annex B) was developed by the American Language-Hearing Association's (ASHA) Division 3: Voice and Voice Disorders (ASHA, 2006) to encourage the standard implementation and documentation of auditory-perceptual VQ evaluation. This clinical and research tool

includes specific phonatory tasks and procedures for voice sample collection and scoring, in order to improve the consistency of clinical evaluation and the exchange of information between clinicians or researchers (Kempster et al., 2009; Nemr et al., 2012; Zraick et al., 2011).

The CAPE-V specifies that the subject whose voice is being evaluated produces three specific phonatory tasks: sustain [a, i], reading aloud of six sentences, and spontaneous speech. This instrument evaluates the subject's performance along all phonatory tasks by rating them in six different vocal parameters labeled and defined as:

- Overall severity: global, integrated impression of voice deviance;
- Roughness: perceived irregularity in the voicing source;
- Breathiness: audible air escape in the voice;
- Strain: perception of excessive vocal effort (hyperfunction);
- Pitch: perceptual correlate of fundamental frequency;
- Loudness: perceptual correlate of sound intensity.

For each vocal attribute, the CAPE-V displays a 100 millimeter line forming a visual-analog scale (VAS) to be used to document each rating. For each vocal attribute, the listener should indicate the degree of perceived deviance from the normal (leftmost portion of the scale) with a tick mark placed along the VAS. A supplement severity indicator is placed beneath each VAS: "MI" or mildly deviant, "MO" or moderately deviant, and "SE" or severely deviant. On the right of each scale there are two letters, "C" and "I", classifying the consistency or intermittent presence of the vocal attribute within or across the phonatory tasks (ASHA, 2006; Kempster et al., 2009; Zraick et al., 2011). The CAPE-V also includes two unlabeled scales that can be used to document other additional perceptual attributes necessary to describe a specific voice, or to note any comments about resonance.

The CAPE-V has been increasingly used both for clinical and research practice (Solomon, Helou & Stojadinovic, 2011). The advantage of the CAPE-V is that its administration and scoring always follows the same procedure, allowing a standardized auditory-perceptual VQ evaluation and documentation across all the vocologists. This instrument can also be applied to normal or dysphonic voices, in adults and children (Jesus et al., 2009b; Jesus et al., 2009a; Karnell et al., 2007; Kelchener et al., 2010;

Mozzanica, Ginocchio, Borghi, Bachmann & Schindler, 2013; Nerm et al., 2012; Nerm, Simões-Zenari, Souza, Hachiya & Tsuji, 2015; Núñez-Batalla, Morato-Galán, García-López & Ávila-Menéndez, 2015; Zraick, et al., 2011). The CAPE-V evaluates more VQ parameters than GRBAS (i.e. pitch and loudness) across several phonatory tasks and allows for the analysis of resonance and two additional not predetermined vocal parameters, enabling a complete voice evaluation and a broader understanding of vocal patterns (Nemr et al., 2012). The CAPE-V VAS has detailed and analytical information about the different vocal parameters assessed, and discriminates small and subtle VQ changes in voice disorders (Nemr et al., 2012). In addition to documenting the severity of the disordered parameters, the CAPE-V also allows for an improved understanding of the anatomical and physiological bases of a voice disorder (Behlau, 2004). The CAPE-V has been translated and adapted into different languages such as: Brazilian Portuguese (BP) (Behlau, 2004), EP (Jesus et al., 2009a), Italian (IT) (Mozzanica et al. 2013), and Spanish (SP) (Núñez-Batalla et al., 2015), promoting an international standardization of auditory-perceptual evaluation across different linguistic and cultural populations.

GRBAS and CAPE-V are widely used by health and/or educational professionals in the voice field (i.e. SLP, ENT, voice teachers) and can be selected depending on specific clinical or research purposes (Nemr et al., 2012). In contrast to GRBAS, CAPE-V has formal administration procedures for voice sample collection and ratings. The definitions of the different vocal parameters are similar in both scales; however, the scales do not use the same exact parameters to characterize VQ. CAPE-V evaluates the same GRBAS parameters with exception of asthenia; it also evaluates two more vocal parameters (i.e. pitch and loudness). In GRBAS, each of the vocal parameters are rated using an ordinal four-point scale, whereas the CAPE-V, uses an interval-level VAS for the same purpose. Based on a comparative analysis between the GRBAS and CAPE-V characteristics, it seems that CAPE-V displays more advantages for the clinical and research purposes, despite demanding more time for administration (see Table 2).

Table 2 – Comparative analysis between GRBAS and CAPE-V instruments.

	GRBAS [1]	**CAPE-V** [2, 3]
Procedures	Not defined	Phonatory tasks, and procedures for voice recording and rating
Phonatory tasks	Not defined	Sustained [a, i]; reading aloud six sentences with specific targets; and spontaneous speech
Vocal parameters	• Grade • Roughness • Breathiness • Strain • Asthenia	• Overall severity • Roughness • Breathiness • Strain --- • Pitch • Loudness
Rating scale	Ordinal scale from 0 to 3	VAS from 0 to 100 mm
Vantages/disadvantages	• No formal administration procedures • Defined terminology • Only assess glottic level --- • Only applicable to dysphonic voice --- • Administration time < 5 minutes • No formalized training • Inter- and intra-rater reliability evidence --- • Assess five vocal parameters • Simple and quick to learn and apply	• Formal administration procedures • Defined terminology • Assess glottic and supra-glottic parameters (i.e. resonance) • Applicable to normal and dysphonic voices • Administration time > 10 minutes • No formalized training • Inter- and intra-rater reliability evidence --- • Assess six vocal parameters • Allow to add additional vocal parameters

[1] Hirano (1981); [2] ASHA (2006); [3] Kempster et al. (2009); VAS – Visual analog scale.

2.2. Validity and Reliability

Any instrument should have strong psychometric characteristics such as acceptable and documented: reliability, validity, specificity and sensibility. The validity of an instrument is important to health outcomes measurement and to the health decision making process that follows (Kelly, O'Malley, Kallen & Ford, 2005). Evidence of validity and reliability are prerequisites to assure the integrity and quality of a measurement instrument (Devon et al., 2007; Kimberlin & Winterstein, 2008). According to the Scientific Advisory Committee of the Medical Outcomes Trust (SACMOT), an instrument may document the health status at a given point in time, distinguish two or more groups, assess any changes over a period of time among groups, and predict future status (Aaronson et al., 2002).

The validity of an instrument is often defined as the degree to which the instrument measures what it purposes to measure (Aaronson et al., 2002; Franic, Bramlett & Bothe, 2005; Kimberlin & Winterstein, 2008; Lohr et al., 1996). In other words, the validity of an instrument relies on its ability to appropriately measure the attributes of the construct under study, through the extent to which the scores or their interpretation are representative of the underlying construct (Devon et al., 2007; Franic et al., 2005; Kimberlin & Winterstein, 2008).

There are different types of validity such as construct, face, content, predictive, concurrent, convergent, and discriminant (Devon et al., 2007). SACMOT determines validity has three aspects: content, construct, and criterion (Aaronson et al., 2002). Content validity reflects the adequacy of the items contained within the instrument to the domain of the instrument (Devon et al., 2007). This type of validity demonstrates if the individual items are a representative sample of the range of items under the construct (Andy, 2009; Cronbach & Meehl, 1955; Kimberlin & Winterstein, 2008). Content validity is generally achieved by using a lay and expert panel that judges the clarity, comprehensiveness, and redundancy of the items and scales of an instrument (Aaronson et al., 2002; Devon et al., 2007; Kimberlin & Winterstein, 2008; Lohr et al., 1996). For example, in a voice evaluation instrument, the content validity can be assured by a panel of experts in voice disorders.

Construct validity is the degree to which an instrument measures the construct under study (Cronbach & Meehl, 1955). This type of validity is supported if the instrument's items are related to theoretical and operational concepts of the construct and supports the measurement of the construct in multiple ways (Aaronson et al., 2002; Devon et al., 2007; Lohr et al., 1996). There are different ways to analyze an instrument's construct validity such as by using contrasted groups, hypothesis testing, factor analysis, and the multitrait-multimethod (MT-MM) approach (Devon et al., 2007). In the contrasted group approach, two groups that are either very similar or complete opposites are sample paired, in order to examine the logical relationship that should exist between the measures or scores on relevant variables (Aaronson et al., 2002; Devon et al., 2007). For example, in a voice evaluation instrument, the construct validity can be determined based a comparative analysis between the results from two different groups such as normal and dysphonic speakers.

Concurrent validity is a type of criterion-related validity, where evidence is showed by the extent to which the scores of the instrument are related to a criterion measure (Aaronson et al., 2002; Lohr et al., 1996). In determining this validity, scores of an instrument are correlated to the scores of another one that measures the same construct in the same subjects (Kimberlin & Winterstein, 2008). This type of validity is confirmed when the scores of two instruments, accepted as theoretically-related and valid for measurement of the same construct, are highly correlated (Aaronson et al., 2002; Devon et al., 2007; Kimberlin & Winterstein, 2008; Lohr et al., 1996). For example, in a voice evaluation instrument, the concurrent validity can be determined by comparing two similar scales such as GRBAS and CAPE-V.

Reliability is the degree to which an instrument is free from random error, or the extent to which obtained scores can be reproduced (Aaronson et al., 2002; Franic et al., 2005; Lohr et al., 1996). There are two classical approaches for examining reliability: internal consistency and reproducibility (e.g. inter-rater reliability and test–retest) (Aaronson et al., 2002; Franic et al., 2005; Lohr et al., 1996); both must be ensured for acceptable reliability of measurement to be established.

Inter-rater reliability determines the equivalence of ratings obtained with an instrument when used by different raters, i.e. it measures the degree of concordance between different raters (Kimberlin & Winterstein, 2008). One way it is estimated is through the intraclass correlation coefficient (ICC), using an analysis of variance to estimate how well ratings from different raters coincide (Cook & Beckman, 2006). For example, in examining the inter-rater reliability of a voice evaluation instrument, a voice sample is rated by different raters to assess their agreement on the different VQ parameters.

Intra-rater reliability or test-retest reliability is the reproducibility or stability measure of an instrument over time (Aaronson et al., 2002; Lohr et al., 1996). This reliability is determined by the administration of the same instrument to the same group of raters at two different times (Aaronson et al., 2002; Devon et al., 2007; Kimberlin & Winterstein, 2008; Lohr et al., 1996). The correlation between the two sets of scores can be determine by using statistical tests such as ICC, pearson correlation, and t test (Devon et al., 2007; Kimberlin & Winterstein, 2008). For example, a voice sample is rated by the same rater at two different times using the same instrument to estimate the intra-rater

agreement for each VQ parameter. For both inter- and intra-rater reliability, the common accepted coefficients thresholds for documentation of acceptable levels are .70 for group comparisons and .90-.95 for individual measurements over time (Aaronson et al., 2002; Lohr et al., 1996).

2.2.1. Validity and Reliability of Auditory-Perceptual Evaluation

A valuable clinical tool must be robust, consistent, and stable (Wuyts et al., 1999). The validity of an instrument also requires that it must be reliable; this is one of the central issues of auditory-perceptual voice evaluation instruments.

Auditory-perceptual voice evaluation is considered to be subjective mainly because it relies on a listener's judgments. The validity and reliability of this type of voice evaluation is influenced by different characteristics of the listeners, the voice stimuli to be judged, and the rating scale used (Bassich & Ludlow, 1986; Bele, 2005; Brinca et al., 2015; Eadie & Baylor, 2006; Eadie et al., 2010; Kreiman & Gerratt, 1998; Kreiman et al., 1990; Kreiman et al., 1993; Kreiman et al., 1992; Maryn & Roy, 2012; Oates, 2009; Sofranko & Prosek, 2012; Wuyts et al., 1999; Zraick et al., 2005). Different studies have pointed out various issues related to inter- and intra-rater reliability in auditory-perceptual evaluation (Bassich & Ludlow, 1986; Kreiman et al., 1990; Kreiman et al., 1993; Maryn & Roy, 2012; Orlikoff, 1999). However, there is some evidence that this variability can be minimized when the factors that influence reliability are identified, and the experimental procedures well designed and controlled (Oates, 2009; Patel, Shrivastav & Edding, 2010).

All listeners' auditory-perceptual judgments of normal and dysphonic VQ can be influenced and susceptible to biases and variability by several factors summarized in Table 3.

Table 3 – Factors that influence listener's auditory-perceptual evaluation.

Factors	Authors
Internal standards	Kreiman, Gerratt & Ito, 2007; Kreiman et al., 2004.
Listener's training	Bassich & Ludlow, 1986; De Bodt, 1997; Eadie & Baylor, 2006; Iwarsson & Peterson, 2012.
Listener's experience and background	Bassich & Ludlow, 1986; Eadie et al., 2010; Helou, Solomon, Henry, Coppit, Howard & Stojadinovic, 2010; Kreiman et al., 1990; Kreiman et al., 1993; Kreiman et al., 1992; Sofranko & Prosek, 2012.
Knowledge of medical diagnosis	Eadie, Sroka, Wright & Merati, 2011a.
Type and length of voice sample/stimulus	Bele, 2005; Brinca et al., 2015; Eadie & Baylor, 2006; Oates, 2009; Zraick et al., 2005.
Degree of pathology	Gerratt et al., 1993.
Task instruction and anchored protocols/stimuli	Awan & Lawson, 2009; Bele, 2005; Eadie & Kapsner-Smith, 2011b; Gerratt et al., 1993.
Type of listening task	Bassich & Ludlow, 1986.
Type of rating scale	Maryn & Roy, 2012; Wuyts et al., 1999.
Number of dimensions rated	Bassich & Ludlow, 1986.

When auditory-perceptual ratings are performed, raters first listen to a voice signal, and then compare it with their internal standards for various properties of voice. These standards are considered to be unstable, because it is thought that internal standards for particular vocal qualities are developed through a listener's unique, previous experiences with voices (Kreiman et al., 1992; McAlliser, Sundberg & Hibi, 1996). These standards can be influenced by the acoustic context in which the voice samples are rated and by a listener's memory of the voice sample last heard (Gerratt et al., 1993; Kreiman et al., 1993; Kreiman et al., 1992). Attention and idiosyncratic sensitivity to certain vocal attributes also are likely to effect a listener's internal standards (Eadie et al., 2010; Kreiman, Gerratt & Berke, 1994). Additional, more random factors belonging to the listeners (e.g. fatigue, lapses, and mistakes) can also influence the intra- and inter-rater reliability of the results (Eadie et al., 2010; Kreiman et al., 1993).

Some findings suggest that clinical training and experience have an important impact on the level of agreement across listeners for VQ (Gerratt et al., 1993; Kreiman et al., 1990; Kreiman et al., 1993). In Portugal, clinicians are considered to be specialists (experts in the area of voice) when they have five or more years of clinical practice with patients with voice disorders. Experienced listeners, especially SLPs who are experts in voice disorders, have been shown to have better inter-rater agreement when compared to inexperienced listeners (De Bodt et al., 1997; Helou et al., 2010; Sofranko & Prosek, 2012; Zraick et al., 2005). Vocal parameters are rated differently by experts in comparison

to naïve listeners, which compromises the reliability of the auditory-perceptual judgments (Kreiman et al., 1994; Kreiman et al., 1990). Expert listeners focus more on breathiness and roughness parameters, and their level of inter-rater agreement is higher on overall severity, breathiness, and roughness (De Bodt et al., 1997; Chan & Yiu, 2006; Iwarson & Peterson, 2012; Karnell et al., 2007; Kreiman & Gerratt, 1998; Webb, Carding, Deary, MacKenzie, Steen & Wilson, 2004).

Listeners have been found to disagree more about slightly and moderately dysphonic voices, than about normal and extremely dysphonic voices (Gerratt et al., 1993; Kreiman et al., 1993). The reliability of ratings increases with the degree of dysphonia (Kreiman & Gerratt, 2011; Law et al., 2012; Rabinov, Kreiman, Gerratt & Bielamowicz, 1995). When training on auditory-perceptual voice evaluation is provided, reliability also increases (Fex, 1992).

Many studies have applied different phonatory tasks (i.e., sustained vowels; reading aloud text; spontaneous speech) to the auditory-perceptual rating of VQ. Spontaneous speech is thought to be the more representative of a person's natural voice (Barsties & De Bodt, 2015; Bele, 2005; McAlliser et al., 1996). The results from this type of phonatory task have shown to be more reliable than sustained vowels (Bele, 2005; Eadie & Doyle, 2005; Law et al., 2012; Zraick et al., 2005). The latter are easier to elicit and allow listeners to judge subtle VQ characteristics without co-articulation effects. However, sustained vowel productions do not incorporate the multidimensional aspects of voice as heard in running speech. When only sustained vowel productions are heard, the auditory-perceptual characteristics seem worse in comparison to connected speech (Zraick et al., 2005).

For a complete auditory-perceptual evaluation of VQ, the selection of the voice samples should combine both phonatory tasks of sustained vowels and running speech. Each tasks enables the clinician to evaluate related, but somewhat different, aspects of VQ. Moreover, improved validity and reliability results when both types of phonatory tasks are included (Law et al., 2012), and their inherent specificities allow a clinician to perform a more comprehensive voice evaluation (Maryn & Roy, 2012).

Type of rating scale may be an important factor in inter-rater reliability (Shrivastav et al., 2005). A VAS appears to allow a finer VQ judgment, offering more detailed information compared to ordinal scales. When a listener is enable to distinguish

a very large number of levels of a VQ parameter, the results will still reflect lack of consistency with some random errors. Inter-rater reliability decreases with an increase of freedom of judgment. (Kreiman et al., 1993; Wuyts et al., 1999). This fact supports the Shrivastav et al. (2005) hypothesis that the variability of inter-rater reliability is related to a listener's use of the scale.

The validity and reliability of auditory-perceptual evaluation results can be increased through the identification and control of the different factors known to influence the auditory-perceptual judgments (see Table 3). Validity and reliability also improve through the systemic use of voice evaluation instruments with predetermined vocal parameters, rating scales, and voice sample testing procedures.

2.2.2. Validity and Reliability of CAPE-V

The CAPE-V is a more recent instrument than GRBAS scale (i.e. 2006 vs 1981, respectively). Several studies have addressed CAPE-V psychometric characteristics – content, construct, and concurrent validity, and inter- and intra-rater reliability (see Table 4). The adequate interpretation of the CAPE-V psychometric characteristics as well as their results should take into account the underlying methodological design and the statistical analysis applied in these studies.

Table 4 – Psychometrics characteristics of CAPE-V.

Study	Instruments	Auditory stimuli	Listeners	Results	Psychometric characteristics analyzed	Limitations of psychometric characteristics
Karnell et al. (2007)	GRBAS CAPE-V V-RQOL IPVI	Dysphonic voice sample: n=34; Phonatory tasks: sustained [a, i], reading aloud six sentences; spontaneous speech.	Raters sample: n=4 SLPs specialized in voice disorders (year of experience NA).	Strong agreement (*r*>.80) between both scales parameters: grade/overall severity, roughness, breathiness, and strain. High intra- and inter-rater reliability (*r*>.80) for CAPE-V overall severity and GRBAS grade parameters.	Concurrent validity. ---------- ---------- Inter-raters reliability; ---------- ---------- ---------- ---------- ---------- ---------- Intra-raters reliability; ----------	• Concurrent validity was performed with one listener (n=1); • Reduced number of listeners (n=4); • Inter-rater reliability was assessed with one parameter: overall severity/grade; • Two voice instruments were applied at the same moment; • Intra-rater reliability was assessed with one parameter: overall severity/grade.
Jesus et al. (2009a)	EP CAPE-V GRBAS	Dysphonic voice sample: n=10; Phonatory tasks: sustained [a, i, u, o], reading aloud six sentences; reading aloud text.	Raters sample: n=2 SLPs specialized (year of experience NA).	High inter-rater reliability and significant correlation for overall severity (ρ=.96, *p*=.00), roughness (ρ=.83, *p*=.01), breathiness (ρ=.99, *p*=.00), and loudness change (k=1.00, *p*=.00). Low inter-rater reliability for pitch (k=.50, *p*=.03). Moderate inter-rater reliability and no statistically significant for strain (ρ=.66, *p*=.08).	Content validity. ---------- ---------- ---------- Concurrent validity. ---------- ---------- ---------- ---------- Inter-rater reliability; ----------	• Sentences proposed do not fulfill all of the CAPE-V original sentences targets. Content validity was not assured; • Two voice instruments were applied at the same moment; • No numerical results about concurrent validity were presented; • Reduced number of listeners (n=2).
Jesus et al. (2009b)	EP CAPE-V GRBAS	Dysphonic voice sample: n=34; Phonatory tasks: sustained [a, i], reading aloud six	Raters sample: n=1 SLPs specialized (year of experience NA).	Good correlation between CAPE-V overall severity and GRBAS grade (ρ=.60, *p*<.005), as well as between CAPE-V and GRBAS breathiness (ρ=.80, *p*<.005). Low correlation between CAPE-V and	Concurrent validity. ----------	• Two voice instruments were applied at the same moment; • Concurrent validity was performed with one listener (n=1); • Correlation between CAPE-V

EP – European Portuguese; BP – Brazilian Portuguese; IT – Italian; SP – Spanish; NA – Not available; LNO – Limitations were not observed.

Table 4 (Cont.) – Psychometrics characteristics of CAPE-V.

Study	Instruments	Auditory stimuli	Listeners	Results	Psychometric characteristics analyzed	Limitations of psychometric characteristics
		sentences; reading aloud text.		GRBAS roughness (ρ=.26, p>.005)		strain and GRBAS strain was not performed.
Kelchener et al. (2010)	CAPE-V	Pediatric dysphonic voice sample: n=50; Phonatory tasks: repeating aloud six sentences.	Raters sample: n=3 SLPs specialized in voice disorders (>7 year of experience).	Moderate to strong inter-rater reliability for overall severity (ICC=67%), roughness (ICC=68%), breathiness (ICC=71%) and pitch (68%) parameters. Low inter-rater reliability for loudness (ICC=63%) and strain (ICC=35%). Intra-rater reliability moderate to strong (ICC=63-87%) for all vocal parameters.	Inter-raters reliability; ---------- ---------- Intra-raters reliability. ---------- ----------	• Reduced number of listeners (n=3); • LNO. ---------- ----------
Zraick et al. (2011)	CAPE-V GRBAS	Normal voice sample: n=22; Dysphonic voice sample: n=37; Phonatory tasks: sustained [a, i], reading aloud six sentences; spontaneous speech.	Raters sample: n=21 SLPs specialized in voice disorders (>5 year of experience).	Strong correlation between the following CAPE-V and GRBAS parameters: overall severity/grade (r=.80), roughness (r=.78), breathiness (r=.78), and strain (r=.77). Inter-rater reliability ranged from high for overall severity (ICC=.76) to low for pitch (ICC=.28). High intra-rater reliability for breathiness (r=.82); good for roughness (r=.77) and loudness (r=.78); and moderate for overall severity (r=.57) and pitch (r=.64). Low for strain (r=.35). Good intra-rater reliability (r>.77) for roughness (14 of 21 raters), breathiness (17 of 21 raters), and loudness	Concurrent validity. Inter-raters reliability; Intra-raters reliability; ---------- ----------	• LNO; • LNO; • Judging sessions with an interval of 48-72 hours. Listeners learning factor could compromised intra-rater reliability. ---------- ----------

EP – European Portuguese; BP – Brazilian Portuguese; IT – Italian; SP – Spanish; NA – Not available; LNO – Limitations were not observed.

Table 4 (Cont.) – Psychometrics characteristics of CAPE-V.

Study	Instruments	Auditory stimuli	Listeners	Results	Psychometric characteristics analyzed	Limitations of psychometric characteristics
				(7 of 21 raters) parameters.		
Nerm et al. (2012)	BP CAPE-V GRBAS	Normal voice sample: n=10; Dysphonic voice sample: n=50; Phonatory tasks: sustained [a, i], reading aloud six sentences; spontaneous speech.	Raters sample: n=3 SLPs specialized in voice disorders (>5 year of experience).	Strong correlation (*r*=.84) between the CAPE-V overall severity and the GRBAS grade parameters. In both scales there was high inter-rater reliability (ICC>.79) for overall severity/grade, roughness, breathiness, and strain. Strong intra-rater reliability (ICC>.93) for CAPE-V overall severity.	Concurrent validity. ---------- ---------- Inter-raters reliability; ---------- ---------- ---------- ---------- ---------- ---------- Intra-raters reliability; ----------	• Concurrent validity was assessed with one parameter: overall severity/grade; • Reduced number of listeners (n=3); • Inter-rater reliability was assessed with CAPE-V parameters: overall severity; roughness, breathiness, and strain; • Intra-rater reliability was assessed with one parameter: overall severity/grade.
Mozzanica et al. (2013)	IT CAPE-V GRBAS	Normal voice sample: n=120; Dysphonic voice sample: n=80; Phonatory tasks: sustained [a, i], reading aloud six sentences; spontaneous speech.	Raters sample: n=3 SLPs specialized in voice disorders (>5 year of experience).	For all six parameters there was significant differences (*p*<.0001) between the control and the dysphonic groups. High correlation (*r*=.92) between CAPE-V overall severity and GRBAS grade parameters; and good correlation between the two scales parameters: roughness (*r*=.84), breathiness (*r*=.87), and strain (*r*=.79). High inter-rater reliability for overall severity (ICC=.92), roughness (ICC=.92), and breathiness (ICC=.90). Good intra-rater reliability for	Content validity; Construct validity; ---------- ---------- ---------- Concurrent validity. Inter-raters reliability; ---------- Intra-raters reliability; ----------	• LNO; • Voice samples were not gender and age balance compromising construct validity; • LNO; • Reduced number of listeners (n=3); • LNO. ----------

EP – European Portuguese; BP – Brazilian Portuguese; IT – Italian; SP – Spanish; NA – Not available; LNO – Limitations were not observed.

Table 4 (Cont.) – Psychometrics characteristics of CAPE-V.

Study	Instruments	Auditory stimuli	Listeners	Results	Psychometric characteristics analyzed	Limitations of psychometric characteristics
				strain (ICC=.89), pitch (ICC=.88), and loudness (ICC=.80).		
Núñez-Batalla et al. (2015)	SP CAPE-V GRBAS	Normal voice sample: n=17; Dysphonic voice sample: n=50; Phonatory tasks: sustained [a, i], reading aloud six sentences; spontaneous speech.	Raters sample: n=2 SLPs specialized in voice disorders (year of experience NA).	High correlation (ICC>.84) between CAPE-V and GRBAS parameters: overall severity/grade, roughness and strain; and moderate (ICC=.61) between CAPE-V and GRBAS breathiness. The sustained vowels task had the highest correlations (ICC>.91) between all the CAPE-V and GRBAS parameters. High inter-rater reliability (ICC>.77) for overall severity, roughness, and breathiness; good (ICC>.65) for strain and pitch; moderate (ICC>.55) for loudness – across all phonatory tasks. High intra-rater reliability (ICC>.85) for all parameters, across all the phonatory tasks.	Content validity. Concurrent validity; ---------- ---------- ---------- ---------- ---------- ---------- ---------- Inter-raters reliability; ---------- Intra-raters reliability; ----------	• LNO; • GRBAS was used to rate sustained vowel task and CAPE-V to rate the three phonatory tasks: sustained vowels, reading aloud, and spontaneous speech. This compromises the concurrent validity; • Reduced number of listeners (n=2); • Intra-rater reliability was determined with one listener; ----------
Nerm et al. (2015)	BP CAPE-V DSI	Normal voice sample: n=42; Dysphonic voice sample: n=24; Phonatory tasks: sustained [a, i], reading aloud six sentences; spontaneous speech.	Raters sample: n=2 SLPs specialized in voice disorders (>5 year of experience).	For all six parameters there was significant differences (p<.0001) between the control and the dysphonic groups.	Construct validity; ----------	• Reduced number of listeners (n=2); • Voice samples were not gender and age balance, compromising construct validity. ----------

EP – European Portuguese; BP – Brazilian Portuguese; IT – Italian; SP – Spanish; NA – Not available; LNO – Limitations were not observed.

CAPE-V's content validity was analyzed into its translation and adaptation in different languages: Brazilian Portuguese (BP), European Portuguese (EP), Italian (IT), and Spanish (SP). This was assured by different professionals, depending on the language translation (summarized in Table 5).

Table 5 – CAPE-V content validity in different languages.

CAPE-V translation	Authors	Content validity review
BP	Behlau (2004)	• Group of SLPs.
EP	Jesus et al. (2009a)	• One speech and hearing scientist; • One linguistic; • Three experienced SLPs.
IT	Mozzanica et al. (2013)	• Consensus of phoniatricians.
SP	Núñez-Batalla et al. (2015)	• One SLP.

The CAPE-V translation into EP was performed by Jesus et al. (2009a). However, the sentences designed for this translation do not accomplish all the original sentences' purposes, nor the phonetic targets determined on the original CAPE-V. Thus, this translation does not guarantee its content validity in relation to the original instrument.

CAPE-V construct validity was reported for the IT and BP versions of CAPE-V (Mozzanica et al., 2013; Nerm et al., 2012). Student's *t*-test was performed to compare the CAPE-V mean scores obtained in normal and dysphonic voice samples for the six CAPE-V parameters. Results revealed significant differences between the groups for all six CAPE-V parameters ($p<.0001$), guaranteeing this psychometric characteristic for IT and BP versions.

CAPE-V concurrent validity was reported in several studies where different methodological procedures were adopted. This may lead to a weaker support for this psychometric characteristic. In the Karnell et al. study (2007), voice ratings were completed using the CAPE-V and GRBAS at the same time, and concurrent validity was estimated with one single listener. Nerm et al. (2015) only provided the correlation between the CAPE-V overall severity and GRBAS grade. In Núñez-Batalla et al. (2015) study, the GRBAS scale was used to only rate vowel production task, while the CAPE-V was used to rate all three CAPE-V phonatory tasks. The concurrent validity was reported based on the ICC results; this lack of consistency in tasks compromises this psychometric characteristic assessment. Differences in statistical analysis can also lead to psychometric problems. The correlation between equivalent CAPE-V and GRBAS parameters was

performed using Spearman's correlation coefficients (Jesus et al., 2009b; Karnell et al., 2007; Nerm et al., 2012; Mozzanica et al., 2013), or multiserial correlation coefficients (Zraick at al., 2011). High correlations (r>.70) were found between the following CAPE-V and GRBAS parameters: overall severity/grade, roughness, breathiness, and strain (see Table 6).

Table 6 – CAPE-V and GRBAS concurrent validity across different studies.

Study	Statistical analysis (>.70)	Vocal parameters			
		Overall severity/grade	Roughness	Breathiness	Strain
Karnell et al. (2007)	*r*	✓	✓	✓	✓
Jesus et al. (2009b)	*ρ*	X	X	✓	X
Zraick et al. (2011)	*r*	✓	✓	✓	✓
Nerm et al. (2012)	*r*	✓	NA	NA	NA
Mozzanica et al. (2013)	*r*	✓	✓	✓	✓

✓ – Correlation higher than .70; X – Correlation lower than .70; NA – Not available.

CAPE-V reliability is a well reported psychometric characteristic. In general, the reliability results can be influenced by differences in the auditory stimuli presented among all CAPE-V studies such as: 1) type of voice sample; 2) voice sample sequence; 3) phonatory tasks; 4) listeners fatigue; 5) listeners training.

In some studies, only dysphonic voices samples were provided to listeners (e.g.: Jesus et al., 2009b; Jesus et al., 2009a; Kelchener et al., 2010), while others provided normal and dysphonic voice samples to be rated (Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015; Zraick et al., 2011). In contrast to the majority of CAPE-V studies, Kelchener et al. (2010) used only pediatric voice samples. The Karnell et al. (2007) was the only investigation that involved voice samples balanced and matched by age and gender.

The voice samples were presented to listeners following the same sequence (Jesus et al., 2009; Jesus et al., 2009a; Kelchner et al., 2010; Mozzanica et al., 2013; Zraick et al., 2011), or following two different randomized sequences (Karnell et al. 2007; Nerm et al., 2012; Núñez-Batalla et al., 2015).

Different phonatory tasks were used in CAPE-V studies. Listeners judged the three CAPE-V phonatory tasks (Karnell et al., 2007; Mozzanica et al., 2013; Nerm et al., 2012; ; Nerm et al., 2015; Núñez-Batalla et al., 2015), while in others they judged some

of them such as: sustained vowels [a, i, u, o], reading aloud CAPE-V sentences, and reading aloud a text (Jesus et al., 2009b; Jesus et al., 2009a); repeating aloud the CAPE-V sentences (Kelchener et al., 2010), or CAPE-V spontaneous speech (Zraick et al., 2011).

Reliability can also be influenced by a listener's fatigue or attention when rating a large number of voice samples. Variability in the total number of voice samples is observed across the several CAPE-V studies (range from 10 to 200) (Jesus et al., 2009b; Jesus et al., 2009a; Karnell et al., 2007; Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015; Zraick et al., 2011).

The training on the rating task was provided to CAPE-V voice rating for four voice samples (Karnell et al., 2007; Zraick et al. 2011) or in one hour of training with some voice samples used as anchor stimuli (Mozzanica et al., 2013).

CAPE-V inter-rater reliability determination has most often been based on a reduced number of listeners (<4), which limited the power of these results (i.e. Jesus et al., 2009a; Karnell et al., 2007; Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015). Zraick et al. (2011) were the only authors who assured a large number of listeners (n=21) for the inter-rater reliability determination. This reliability has been estimated based on different statistical analysis such as ICC determination (Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015; Zraick et al., 2011), and Spearman's correlation coefficient (Jesus et al., 2009a; Karnell et al., 2007). Inter-rater reliability was high (>.70) for the following CAPE-V parameters: overall severity, roughness, breathiness, strain, pitch, and loudness (resumed in Table 7).

Table 7 – CAPE-V inter-rater reliability across different studies.

Study	Statistical analysis (>.70)	Vocal parameters					
		Overall severity	Roughness	Breathiness	Strain	Pitch	Loudness
Karnell et al. (2007)	*r*	✓	NA	NA	NA	NA	NA
Jesus et al. (2009a)	*ρ*	✓	✓	✓	X	X	✓
Kelchener et al. (2010)	ICC	X	X	✓	X	X	X
Zraick et al. (2011)	ICC	✓	X	X	X	X	X
Nerm et al. (2012)	ICC	✓	✓	✓	✓	NA	NA
Mozzanica et al. (2013)	ICC	✓	✓	✓	✓	✓	✓
Núñez-Batalla et al. (2015)	ICC	✓	✓	✓	X	✓	X

✓ – Correlation higher than .70; X – Correlation lower than .70; NA – Not available.

CAPE-V intra-rater reliability was also studied. The common methodological limitation to this reliability is related to a learning factor that can occur when repeated voice samples are rated by listeners. Intra-rater reliability results are likely influenced by: 1) rating session characteristics; 2) sequence and number of repeated voice samples presented to listeners; 3) number of vocal parameters assessed; and 4) statistical analysis.

In most of the studies, repeated voice sample rating were separated by a one week of interval (Karnell et al., 2007; Kelchener et al. 2010; Mozzanica et al., 2013; Núñez-Batalla et al., 2015). In the Nerm et al. (2012) and Zraick et al. (2011) studies, repeated voice samples were presented during the same listening session. Some differences, possibly related to the number of repeated voice samples, were also found across the several studies. Listeners judged all voice samples twice (Karnell et al., 2007; Mozzanica et al., 2013), while in others they rated a subset of total voice samples (Kelchener et al., 2010; Nerm et al., 2012; Núñez-Batalla et al., 2015; Zraick et al., 2011).

The number of vocal parameters assessed in intra-rater reliability varied between one (Karnell et al., 2007; Nerm et al., 2012) and six (Kelchener et al., 2010; Mozzanica et al., 2013; Núñez-Batalla et al., 2015; Zraick et al., 2011).

Intra-rater reliability results were based on different statistical analyzes such as ICC (i.e. Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015), Spearman's correlation coefficients (Karnell et al., 2007), and Pearson correlation coefficients (Zraick et al., 2011). High intra-rater reliability (>.70) was reported for CAPE-V parameters, across several studies (summarized in Table 8).

Table 8 – CAPE-V intra-rater reliability across different studies.

Study	Statistical analysis (>.70)	Vocal parameters					
		Overall severity	Roughness	Breathiness	Strain	Pitch	Loudness
Karnell et al. (2007)	*r*	✓	NA	NA	NA	NA	NA
Kelchener et al. (2010)	ICC	✓	✓	✓	X	✓	✓
Zraick et al. (2011)	*r*	X	✓	✓	X	✓	X
Nerm et al. (2012)	ICC	✓	NA	NA	NA	NA	NA
Mozzanica et al. (2013)	ICC	✓	✓	✓	✓	✓	✓
Núñez-Batalla et al. (2015)	ICC	✓	✓	✓	✓	✓	✓

✓ – Correlation higher than .70; X – Correlation lower than .70; NA – Not available.

Although all the methodological differences across the several CAPE-V studies (Jesus et al., 2009b; Jesus et al., 2009a; Karnell et al., 2007; Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Nerm et al., 2015; Núñez-Batalla et al., 2015; Zraick et al., 2011), their results support the validity and reliability of CAPE-V for the both clinical and research auditory-perceptual voice evaluation purposes.

2.3. Definition of the Problem

The research results of different CAPE-V studies emphasize the validity and reliability of this instrument when applied to both clinical and research fields. However, the results reported in the previous chapter could have been influenced by methodological limitations (see Table 4), e.g., number of listeners, selection of dysphonic voice samples only, and number and type of phonatory tasks.

A comparative analysis was performed and revealed some disparities among the American English (AE), EP and BP CAPE-V versions (see Table 9):

- AE 1st edition (ASHA, 2006);
- BP version (Behlau, 2004)
- EP 1st version (Jesus et al., 2009a);
- AE 2nd edition (Kempster et al., 2009).

As recommended by SACMOT (Aaronson et al., 2002), any health status and quality-of-life assessment instrument must be valid and reliable. The AE CAPE-V (ASHA, 2006; Kempster et al., 2009) cannot be applied to EP or any other language because of the linguistics differences between these languages. However, sentences must target the same phonetic features (i.e.: vowel production; soft glottal attack; all voiced phonemes; vowel initiated words; nasal consonants; no nasal consonants) in both languages. In the two AE CAPE-V versions (ASHA, 2006; Kempster et al., 2009) there are slight differences in the sentences targets (i.e.: production of every vowel in English vs coarticulatory influence of three vowels; easy onset with [h] vs soft glottal attacks and voiceless to voiced transition; weighted with voiceless plosive sounds vs contains no nasal consonants). BP version of CAPE-V (Behlau, 2004) can-not be applied to EP as well, due to phonetic differences between these two languages. Specific linguistic characteristics do not guarantee content validity when BP CAPE-V is applied to EP linguistic context, because some of the sentences targets established on the original CAPE-V sentences are missing:

- Sentence A does not include the EP oral vowel [a];
- Sentence C does not include the voiced EP phonemes (i.e. [b, d, g, ʒ, m, n, ɲ, l, ʁ]);
- Sentence D does not include the EP hard attack vowels (i.e. [a, i, u]);

- Sentence E does not include the EP nasal vowels (i.e. [ĩ, õ, ẽ]);
- Sentence F contains the EP nasal phoneme [ẽ]

A psychometric analysis of the EP version of CAPE-V (Jesus et al., 2009a) (Annex C), was performed (see Table 10). Content validity was not achieved because the sentences proposed did not mirror the CAPE-V original sentences purposes nor the phonetic targets (see Table 9) e.g.:

- Sentence B included one word which begin with voiced phoneme [ɗa];
- Sentence C did not include the voiced EP phonemes (i.e. [b, g, ʒ, ɲ, m, l]);
- Sentence D did not include the EP hard attack vowels (i.e. [a, i, u]);
- Sentence E did not include the EP nasal vowels (i.e. [ĩ,õ,ũ]) and the consonant [n];
- Sentence F contained the nasal phoneme [m] and the voiced plosive [g], which were not targets.

EP CAPE-V construct validity was not assessed and could not be guaranteed because only disordered voices samples were included. Inter-rater reliability was measured based on the scores of two raters, which limited the degree to which an instrument can be found to be free from random error. Intra-rater reliability was not evaluated in the EP version of CAPE-V (Jesus et al., 2009b; Jesus et al., 2009a). Therefore, a second translation of CAPE-V into EP was needed to be developed, where content, construct, and concurrent validity was supported as well as inter- and intra-rater reliability (see Table 10).

Table 9 – Comparative analysis among four CAPE-V versions.

	AE CAPE-V – 1st Ed. (ASHA, 2006)	BP CAPE-V (Behlau, 2006)	EP CAPE-V (Jesus et al., 2009a)	AE CAPE-V – 2nd Ed. (Kempster et al., 2009)
Sentence A	**Target:** "Provides production of every vowel in the **English language**". "The blue spot is on the key again"	**Target:** "Provides production of every vowel in the **BP**". "Érica tomou suco de pêra e amora" [ˈɛrikɐtumoˈsukuɖɨperɐjɐmɔrɐ]	**Target:** "Provide production of every oral vowel in **EP**". "A Marta e o avô vivem naquele casarão rosa velho" [ɐ̃ˈmartɐjuɐˈvoˈvivɐ̃jnɐkeliɬkɐˈrɐ̃wˈʁɔzɐˈvɛʌu]	**Target:** "Examine **coarticulatory influence of three vowels [a, i, u]**. "The blue spot is on the key again"
Analysis	NA to EP.	Not include EP oral vowel [a]	Include all the EP oral vowels.	NA to EP.
Sentence B	**Target:** "Emphasizes **easy onset with the [h]**". "How hard did he hit him?"	**Target:** "Emphasizes **easy onset with the [s]** "Sónia sabe sambar sozinha" [ˈsɔnjɐˈsabɨsɐ̃ˈbarsɔˈziɲɐ]	**Target:** "**Easy onset with [s]**". "Sofia saiu cedo da sala" [ˈsufiɐsɐˈiwˈseɖuɖɐˈsalɐ]	**Target:** "**Assess soft glottal attacks** and voiceless to voiced transition". "How hard did he hit him?"
Analysis	NA to EP	It has all words begin with easy onset [s].	It has one word that doesn't begin with easy onset [s], i.e. [ɖɐ] which begins with voiced phoneme.	NA to EP
Sentence C	**Target:** "**All voiced**". "We were away a year ago"	**Target:** "**Voiced segments**". "Olha lá o avião azul" [ˈɔʌɐˈlauɐviɐ̃wɐˈzuɬ]	**Target:** "**Only voiced** phonemes". "A asa do avião andava avariada" [ɐˈazɐɖuɐviˈɐ̃wɐ̃ɖavɐvɐriˈaɖɐ]	**Target:** "**Features all voiced phonemes** and provides a context to judge possible voiced stoppages/spasms and one's ability to link from one word to another". "We were away a year ago"
Analysis	NA to EP	It has only voiced phonemes. However it does not include the voiced EP phonemes [b, d, g, ʒ, m, n, ɲ , l, ʁ].	It has only voiced phonemes. However it not include the voiced EP phonemes [b, g, ʒ, ɲ, m l].	NA to EP

AE – American English; BP - Brazilian Portuguese; EP – European Portuguese; NA – not applicable.

Table 9 (Cont.) – Comparative analysis among four CAPE-V versions.

	AE CAPE-V – 1st Ed. (ASHA, 2006)	BP CAPE-V (Behlau, 2006)	EP CAPE-V (Jesus et al., 2009a)	AE CAPE-V – 2nd Ed. (Kempster et al., 2009)
Sentence D	**Target:** "Elicit **hard vocal attacks**". "We eat eggs every Easter"	**Target:** "Elicit **hard vocal attacks**". "Agora é hora de acabar" [ɐˈɣɔɾɐˈɛˈɔɾɐɗɐkɐˈβa]	**Target:** "**Hard glottal attack**". "Agora é hora de acabar" [ɐˈɣɔɾɐˈɛˈɔɾɐɗɐkɐˈβar]	**Target:** "Includes several **vowel-initiated words that may provoke hard glottal attacks** and provides the opportunity to assess whether these occur". "We eat eggs every Easter"
Analysis	NA to EP	It does not include the hard attack EP vowels [a, i, u].	It does not include the hard attack EP vowels [a, i, u].	NA to EP
Sentence E	**Target:** "**Incorporates nasal sounds**". "My mama makes lemon jam"	**Target:** "Assess **nasal sounds** emission". "Minha mãe namorou um anjo" [ˈmiɲɐˈmɐ̃jnɐmuˈroũɐ̃ʒu]	**Target:** "**Nasal phonemes**". "A minha mãe mandou-me embora" [ɐˈmiɲɐˈmɐ̃jmɐ̃domɐ̃bɔɾɐ]	**Target:** "**Includes numerous nasal consonants**, thus providing an opportunity to assess hyponasality and possible stimulability for resonant voice therapy". "My mama makes lemon jam"
Analysis	NA to EP	It does not include the nasal EP vowels [ĩ, õ, ẽ].	It does not include the nasal EP vowels [ĩ, õ, ũ] and the consonant [ɲ].	NA to EP
Sentence F	**Target:** "**Weighed with voiceless plosive** sounds". "Peter will keep at the peak"	**Target:** "**With voiceless plosive** sounds". "Papai trouxe pipoca quente" [pɐˈpajˈtrosɬpiˈpɔkɐˈkẽtɬ]	**Target:** "**Voiceless stops**". "O Tiago comeu quatro peras" [uˈtiaɣukuˈmewˈkwatruˈpeɾɐʃ]	**Target:** "**Contains no nasal consonants** and provides a useful context for assessing intraoral pressure and possible hypernasality or nasal air emission". "Peter will keep at the peak"
Analysis	NA to EP	It contains the nasal phoneme [ẽ] which is not a target for voiceless plosives.	It contains a nasal phoneme [m] and a voiced plosive [g] which is not a target for voiceless plosives.	NA to EP

AE – American English; BP - Brazilian Portuguese; EP – European Portuguese; NA – not applicable.

Table 10 – Analysis and proposal for study the validity and reliability of the II EP CAPE-V.

		Study		
Authors		Jesus et al. (2009a)	Jesus et al. (2009b)	Present study
Instruments		1st EP CAPE-V GRBAS	1st EP CAPE-V GRBAS	II EP CAPE-V GRBAS
Psychometric characteristics	Validity			
	1. Content	The proposed sentences of the 1st version EP CAPE-V do not fulfill all the phonetic requirements of AE CAPE-V.	Not performed.	6 new sentences reviewed by an EP Linguist were proposed to ensure that all phonetics targets of AE CAPE-V 2nd edition (Kempster et al., 2009).
	2. Construct	Not performed.	Not performed.	Contrasted groups approach was used between control and dysphonic group, to observe if there were significant differences (α=.05) in all the auditory-perceptual parameters.
	3. Concurrent	No numerical results about concurrent validity were presented.	Good correlation between GRBAS and EP CAPE-V's overall severity (ρ=.60, p<.005) and breathiness (ρ =.80, p<.005).	Multi-serial correlations between GRBAS and II EP CAPE-V parameters was used.
	Reliability			
	1. Inter-rater reliability	High inter-rater reliability between two listeners for overall severity (ρ=.964, p=.000), roughness (ρ=.991, p=.000), breathiness (ρ=.991, p=.000) and loudness parameters (k=1.000, p=.000).	Not performed.	Inter-raters reliability was performed with 14 listeners.
	2. Intra-rater reliability	Not performed.	Not performed.	Intra-raters reliability was performed with test-retest on 6 repeated voice samples.

EP – European Portuguese; AE – American English.

2.4. Statement of the Purpose

The purpose of the present study was to develop a valid and reliable EP version of the AE 2nd edition of CAPE-V (Kempster et al., 2009) based on psychometric characteristics recommended by SACMOT (Aaronson et al., 2002). This will result in a 2nd EP version of CAPE-V (II EP CAPE-V).

In the present study, content validity of this instrument was supported by the adaptation of the phonatory tasks of sentences, and spontaneous speech. In order to fulfill the requirements stated in the AE 2nd edition of CAPE-V (Kempster et al., 2009), six new sentences were proposed to correspond with the original sentences' targets:

1. Oral and nasal vowel coarticulatory productions;
2. Soft glottal attacks production;
3. Inclusion of only voiced phonemes;
4. Hard glottal attacks production;
5. Strong nasal environment;
6. Inclusion of many voiceless plosives.

Sentences were conceptualized and adapted to the EP linguistic and cultural contexts. They were reviewed by a Portuguese Linguist. For spontaneous speech elicitation, using the prompt "Tell me about the place where you grew up" was proposed, as was suggested on the standardized procedures of CAPE-V (Zraick et al., 2011).

Construct validity was supported by using a contrast groups approach between the control group (CG) and the dysphonic group (DG) in all the CAPE-V vocal parameters.

Concurrent validity was measured by multi-serial correlation between II EP CAPE-V and GRBAS parameters (i.e.: overall severity/grade; roughness; breathiness; and strain). This correlation coefficient is appropriate to use when one variable is interval (CAPE-V) and the other is ordinal (GRBAS).

Reliability of the II EP CAPE-V was estimated by measuring the inter-rater reliability (degree of agreement between listeners) and intra-rater reliability (test-retest).

In order to better study the validity and reliability of II EP CAPE-V, a larger number of listeners (n=14) was used and speakers were matched by age and gender.

2.5. Research Questions

The main goal of this study was to determine the validity and reliability of II EP version of CAPE-V. The research questions addressed were:

1. Validity of II EP version of CAPE-V:
   1.1. Was content validity supported in the II EP version of CAPE-V?
   1.2. Was there a significant difference in VQ between normal and dysphonic voice samples detected by the listeners in all the VQ parameters?
   1.3. What was the correlation between ratings using the CAPE-V and GRBAS auditory-perceptual parameters?
2. Reliability of II EP version of CAPE-V:
   2.1. What was the level of agreement between different listeners in all CAPE-V parameters (inter-rater reliability)?
   2.2. What was the level of agreement among repeated voice sample rated by the same listener (intra-rater reliability)?

2.6. Hypothesis

The following hypothesis are stated were tested:

1. H_0: Auditory-perceptual ratings of the CG were not significantly different when compared to the DG.
   H_1: Auditory-perceptual ratings of the CG were significantly different when compared to the DG.
2. H_0: CAPE-V auditory-perceptual parameters were not highly correlated with the GRBAS auditory-perceptual parameters.
   H_1: CAPE-V auditory-perceptual parameters were highly correlated with the GRBAS auditory-perceptual parameters.
3. H_0: Listeners were found not to have a high level of agreement (reliable) when rating the CAPE-V auditory-perceptual parameters.
   H_1: Listeners were found to have a high level of agreement (reliable) when rating the CAPE-V auditory-perceptual parameters.
4. H_0: Listeners were found not to have a high level of agreement (reliable) when rating the auditory-perceptual parameters of repeated voice samples.

H_1: Listeners were found to have a high level of agreement (reliable) when rating the auditory-perceptual parameters of the repeated voice samples.

III. METHODOLOGY

3.1. Research Design

The data collection for this study was performed during two listening sessions, always following the same procedure; therefore this is considered a transversal study (Groove & Shoyer, 2000; McBurney & White, 2007; Vilelas, 2009). This investigation was also an observational study, because of the observations made (i.e. the ratings) and the analyses of the auditory-perceptual parameters of II EP CAPE-V on the dysphonic and normal voices (McBurney & White, 2007). This study compared normal and dysphonic voices on auditory-perceptual parameters, in order to characterize them (Fortin, 1996; Vilelas, 2009); therefore, it can also be characterized as a comparative study.

The dependent variables of this study were the auditory-perceptual parameters measured with the CAPE-V: overall severity, roughness, breathiness, strain, pitch, and loudness; and GRBAS: grade, rough, breathy, asthenic, and strained. These variables were measured using both a VAS and an ordinal scale. The CAPE-V parameters were quantitative metric variables and GRBAS scale parameters were quantitative ordinal. The independent variables were gender and age of the speaker, and the category of the voices as normal or dysphonic. These variables were classified as qualitative nominal for gender and voice category, and quantitative metric for the age variable.

3.2. Subjects

In this study there were two different subjects: speakers and listeners.

3.2.1. Speakers

The sample of a speaker was obtained using a nonrandom convenience sample, whose selection was based on the practical reason of presence or absence a voice disorder with dysphonia, confirmed by a laryngoscopy (McBurney & White, 2007). The speaker subjects were recruited from the ENT appointment at Hospital da Luz, and underwent a clinical laryngeal evaluation, including a direct laryngoscopy conducted by an ENT

Specialist from the Department. All the subjects signed the informed consent (Appendix A) approved by the Ethics Committee of Hospital da Luz.

Twenty subjects participated in this study: 10 males (mean age=45) and 10 females (mean age=43). Subjects were divided into two different groups: control group (CG=10) and the dysphonic group (DG=10) (see Table 11), matched for age and gender.

Table 11 – Speakers sample size by groups.

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

The selection of all subjects was based on the direct laryngoscopy results, following the scheme described in the Classification Manual for Voice Disorders – I (Verdolini et al., 2006). DG included subjects with different dysphonia etiologies classified in four different groups: structural (n=5), inflammatory (n=1), neurological (n=2), and other disorders (n=2) (see Table 12) (Appendix B).

CG included 10 normal-speakers who fit the following inclusion criteria: 1) no organic or functional laryngeal disorder confirmed by direct laryngoscopy; 2) native EP speaker; 3) over 18 years old; 4) literacy abilities; 5) no voice disorder identified by an SLP using II EP CAPE-V.

DG included 10 subjects who fill the inclusion criteria of: 1) presence of organic or functional laryngeal disorder confirmed by direct laryngoscopy; 2) native EP speaker; 3) over 18 years old; 4) literacy abilities; 5) voice disorder identified by an SLP using CAPE-V. Exclusion criterion were: 1) history of cognitive, or speech and language disorders; 2) allergy, vocal complaints, and/or breathing problems on the day of voice recording.

Table 12 – Distribution of DG according to dysphonia etiology classification.

Classification of voice disorder	n	Gender	n
Structural Pathologies	5	M F	2 3
Inflammatory Conditions	1	M F	1 0
Neurological Disorders	2	M F	1 1
Other disorders	2	M F	1 1
Total	10		10

M – Male; F – Female.

3.2.2. Listeners

Fourteen SLPs who specialize in voice disorders were recruited as listeners; this was also a nonrandom convenience sample. The selection as a listener was based on professional experience with voice disorders (McBurney & White, 2007). Two men (mean age=28) and twelve women (mean age=38) participated as listeners with an average of 11 years of clinical voice experience (see Table 13 and 14).

SLP's signed an informed consent (Appendix C) approved by the Ethics Committee of Hospital da Luz. Inclusion criteria were: 1) more than 5 years of voice clinical experience; 2) caseload of voice patients seen weekly; 3) bilateral normal hearing limits for speech production; 4) knowledge of the CAPE-V instrument for the evaluation of VQ; 5) knowledge and use of the GRBAS scale; 6) native EP speaker. Exclusion criterion was: 1) history of cognitive, or speech and language disorder.

Table 13 – Distribution of listener's subjects by age.

Age (yrs.)	n	Gender	n
17-29	6	M F	2 4
30-39	4	M F	0 4
40-49	1	M F	0 1
50-59	2	M F	0 2
> 60	1	M F	0 1
Total	14		14

M – Male; F – Female.

Table 14 – Distribution of listener's subjects by years of experience.

Year of experience	n	Gender	n
3-5	2	M	0
		F	2
6-10	7	M	2
		F	5
10-20	2	M	0
		F	2
>20	3	M	0
		F	3
Total	14		14

M – Male; F – Female.

3.3. Equipment

Voice samples were captured with headset microphone (PYLE PMEMI), electret condenser, omnidirectional with frequency response 20Hz- 20KHz and sensitivity -44dB± 3dB, and recorded on a portable digital recorder (TASCAM DR-05), 16 bits, mono, with a sample frequency of 44100 Hz. Ambient noise was always below 50 dB, confirmed by a digital sound level meter, model Rolls SLM305. Equipment was always tested and calibrated with a reference pure tone of 500 Hz, confirmed by acoustic analysis. This tone was recorded at the beginning of each recording day.

For voice sample analysis, 14 listeners used the II EP CAPE-V (Appendix D) and GRBAS scale (Annex A).

3.4. Instruments

3.4.1. II EP CAPE-V

II EP CAPE-V (Appendix D) is an instrument for auditory-perceptual voice evaluation with determined voice data collection and scoring procedures. Voice sample was composed by three phonatory tasks: sustained [a, i] three times for 3-5 seconds, reading aloud six sentences, and 20 seconds of spontaneous speech in response to the prompt "Tell me about the place where you grew up".

Based on listening to the three phonatory tasks, the listener judged VQ on six different vocal parameters: 1) overall severity, 2) roughness, 3) breathiness, 4) strain, 5)

pitch, and 6) loudness. Resonance and two additional perceptual attributes could also be judged.

Each VQ parameter was judged using a VAS of 100 millimeter displayed in front of it. The degree of deviance is marked using a tick mark on the scale. The leftmost portion of the line reflects the normal voice or the nonexistence of the VQ parameter being judged. The right end of the scale reflects the most extreme deviance a listener might perceived. Below the VAS line, general regions are displayed as supplement severity indicator. "MI" refers to mildly deviant, "MO" moderately deviant and "SE" severely deviant. At the right of each scale there are two letters, "C" and "I" that represent "consistent" ("C") or "intermittent" (I) presence of a particular vocal parameter within or across phonatory tasks. The judgement of consistency or intermittency was indicated by circling either "C" or "I".

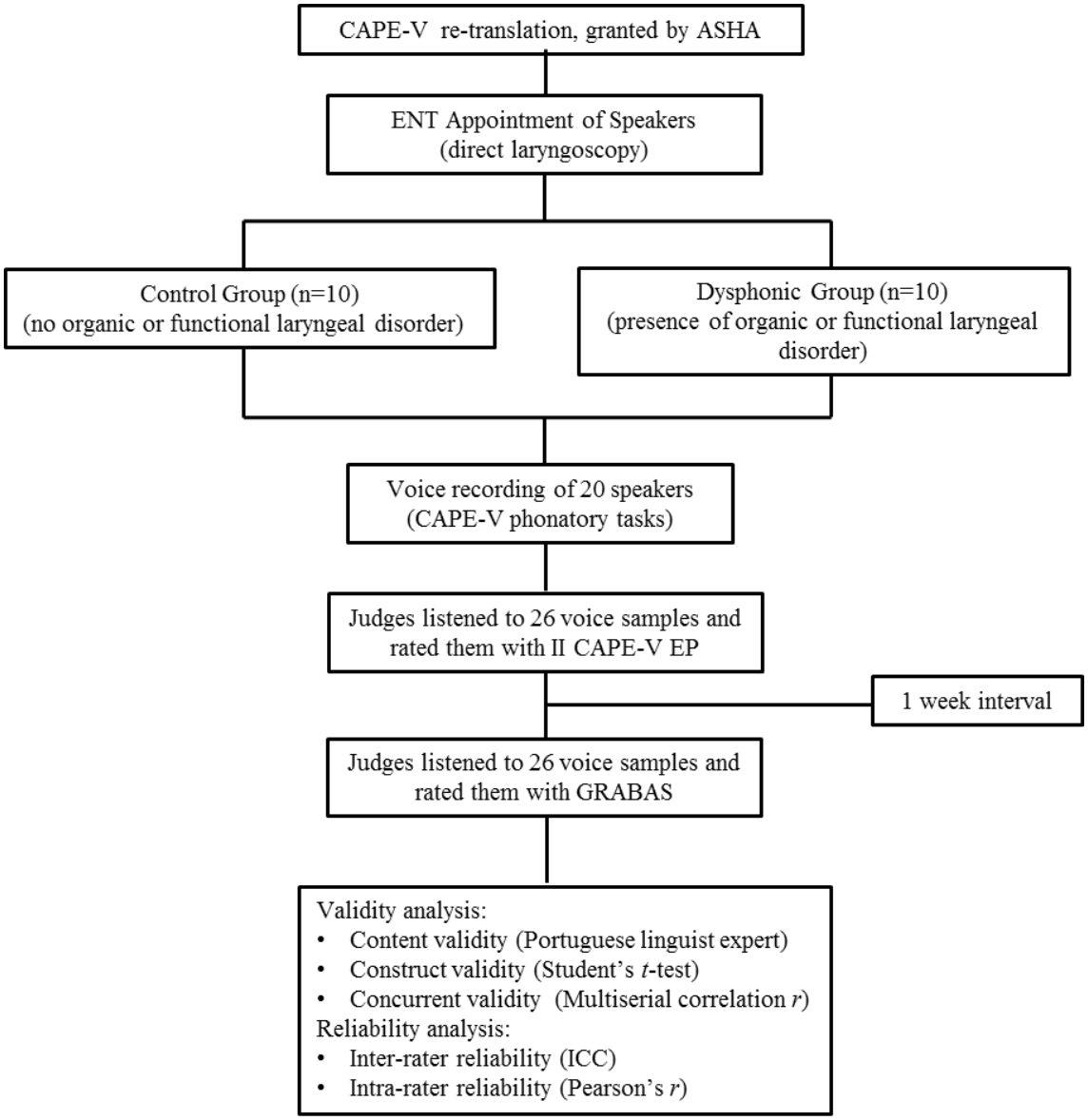
3.4.2. GRBAS

GRBAS scale (Annex A) does not offer a protocol with any specific procedures for voice sample collection, documentation, or evaluation. This scale allows the evaluation of the following vocal parameters: Grade (G), rough (R), breathy (B), asthenic (A), and strained (S). Each parameter is evaluated in a four-point scale from 0 to 3. "0" classification means normal, "1" slight, "2" moderate, and "3" extreme.

3.5. Procedures

This study involved the following different procedures: CAPE-V translation, voice sample recording, and voice samples listening and scoring by listeners using the II EP CAPE-V and GRBAS scales. See below (Figure 1) all the steps taken during this study.

Figure 1 - Present study procedures.



3.5.1. CAPE-V translation

A critical analysis of the 1st translated version of the CAPE-V into EP (Jesus et al., 2009a) (Annex C) was performed. This analysis revealed that proposed sentences did not achieve all the targets intended by the original version of CAPE-V (ASHA, 2006) (Table 9). For the six sentence target established in the AE CAPE-V version (ASHA, 2006; Kempster et al., 2009), none were accurate:

Sentence A "*A Marta e o avô vivem naquele casarão rosa velho*" (*Marta and grandfather live in that old big pink house*) included all EP oral vowels but none of the nasal ones. This compromised the complete assessment of all the EP vowels

coarticulation (oral and nasal). Sentence B "*Sofia saiu cedo da sala*" (*Sofia left the room early*) had one word that did not begin with easy onset [s]. The word [ɗɐ] begins with a voiced phoneme [ɗ]. Sentence C "*A asa do avião andava avariada*" (*The airplane wings was broken*) did have only EP voiced phonemes but was missing several phonemes [b, g, ʒ, ɲ, m, l] . Sentence D "*Agora é hora de acabar*" (*Now it is time to finish*) did not include all the hard glottal attack EP vowels, missing [a, i, u]. Sentence E "*A minha mãe mandou-me embora*" (*My mother sent me away*) did not include EP nasal vowels [ĩ, õ, ũ] as well as the consonant [n]. Sentence F "*O Tiago comeu quarto pêras*" (*Tiago ate four pears*) contained all the EP voiceless plosives sounds. However it also contained a nasal phoneme [m] and a voiced plosive sound [g], which were not a target which altered the intended phonetic context.

Based on the above analyses, six new sentences were designed in order to fulfill the sentence target requirements established under the AE 2nd edition of CAPE-V (Kempster *et al.*, 2009). The original sentence targets were: 1) oral and nasal vowel coarticulatory productions; 2) soft glottal attacks production; 3) inclusion of only voiced phonemes; 4) hard glottal attacks production; 5) strong nasal environment; and 6) inclusion of many voiceless plosives. They were conceptualized and adapted to the EP linguistic and cultural context and were reviewed by a Portuguese linguist expert (see Table 15).

The following sentences were proposed in order to achieve each target objectives:

- Sentence A "*Num domingo esteve sol e fui com o avô António à explanada Évora comer uma empada*" (*On Sunday it was sunny and I went with grand-father António to the terrace of the "Évora"cafe to eat a pie*) to examine the coarticulatory production of all oral and nasal EP vowels;
- Sentence B "*Segundo Simão, só Samuel sabe*" (*According to Simão, only Samuel knows*) to assess soft glottal attacks voiceless to voiced transition through a sentence that only contains words that emphasize the easy onset with [s]. In EP words with easy onset [h] do not exist; therefore, it was proposed to substitute those words for words beginning with easy onset [s].

- Sentence C "*A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa*" (*Zé, Gabriel's mother, gave him an orange cake and old wine from Runa*) to produce all EP voiced phonemes which allows for assessment of possible voiced stoppages/spasms. This new sentence includes all the EP voiced phonemes.
- Sentence D "*É hora da Urraca ir à caça*" (*It is time for Urraca to go hunting*) has words beginning with vowels that elicit glottal attack. This sentence includes all the EP vowels produced in hard glottal attack.
- Sentence E "*Onde eu brinco há um ninho de andorinhas encostado ao muro*" (*Where I play, there is a swallow's nest next to the wall*) includes all the EP nasal vowels and consonants providing the opportunity to assess hyponasality and possible stimulability for resonant voice therapy. In this new sentence all the EP nasal vowels and consonants were included.
- Sentence F "*A Kika tapou a tua capa preta*" (*Kika covered your black cape*) is weighted with voiceless plosive sounds and without any nasal sounds to provide a useful context for assessing intraoral pressure and possible hypernasality or nasal air emission. This sentence has three of each of the voiceless plosive sounds [p, t, k].

After modifying the sentences to address these content validity errors on the 1st EP CAPE-V translation (Jesus et al., 2009a), permission from ASHA was requested to construct a 2nd EP version of the 2nd AE edition of CAPE-V (Kempster et al., 2009). This request included the proposal of six new sentences adapted to EP and to use the prompt "*Tell me about the place where you grew up*" for elicitation of spontaneous speech, as was recommended on the standardized procedures of the CAPE-V (Zraick et al., 2011). ASHA granted non-exclusive permission to translate and reprint the CAPE-V instrument, descriptions, and instructions into EP for use in this research project (Annex D).

Table 15 – Critical analysis of CAPE-V versions and proposal of new sentences.

	CAPE-V				
	AE CAPE-V – 1st Ed. (ASHA, 2006)	BP CAPE-V (Behlau, 2004)	EP CAPE-V (Jesus et al., 2009a)	AE CAPE-V – 2nd Ed. (Kempster et al., 2009)	II EP CAPE-V (present study)
Sentence A	**Target:** "Provides production of every vowel in the English language". "The blue spot is on the key again"	**Target:** "Provides production of every vowel in the BP". "Érica tomou suco de pêra e amora" [ˈɛrikɐtumoˈsukuɖɨperɐjɐmɔrɐ] Not include EP oral vowel [a].	**Target:** "Provide production of every oral vowel in EP". "A Marta e o avô vivem naquele casarão rosa velho" [ɐ̃ˈmartɐjuɐˈvoˈvivɐ̃ʃnɐkelɨkɐˈrɐ̃wˈʁɔzɐˈvɛʌu] Include all the EP oral vowels.	**Target:** "Examine coarticulatory influence of three vowels [a, i, u]". "The blue spot is on the key again"	**Target:** Examine coarticulatory influence of all the oral and nasal EP vowels. "Num domingo esteve sol e fui com o avô António à explanada Évora comer uma empada" [nũˈdumĩguˈʃtevɨˈsɔɬiˈfujkõuɐˈvoɐ̃ˈtɔnjwaʃplɐˈnaɖɐˈɛvurɛkuˈmerumɐɐ̃ˈpaɖɐ] Sentence has all EP oral and nasal vowels.
Sentence B	**Target:** "Emphasizes easy onset with the [h]". "How hard did he hit him?"	**Target:** "Emphasizes easy onset with the [s] "Sónia sabe sambar sozinha" [ˈsɔnjɐˈsabɨsɐ̃ˈbarsɔˈziɲɐ] It has all words begin with easy onset [s].	**Target:** "Easy onset with [s]". "Sofia saiu cedo da sala" [ˈsufiɐsɐˈiwˈseɖuɖɐˈsalɐ] It has one word that does not begin with easy onset [s], i.e. [ɖɐ] which begins with voiced phoneme.	**Target:** "Assess soft glottal attacks and voiceless to voiced transition". "How hard did he hit him?"	**Target:** Asses soft glottal attacks voiceless to voiced transition through a sentence that only contains words that emphasize the easy onset with [s]. "Segundo Simão, só Samuel sabe" [sɨˈɣũdusiˈmɐ̃wˈsɔsɐˈmuɐɬˈsaβɨ] In EP the are no words with easy onset [h] but there is words beginning with easy onset [s].

Table 15 (Cont.) – Critical analysis of CAPE-V versions and proposal of new sentences.

	CAPE-V				
	AE CAPE-V – 1st Ed. (ASHA, 2006)	BP CAPE-V (Behlau, 2004)	EP CAPE-V (Jesus et al., 2009a)	AE CAPE-V – 2nd Ed. (Kempster et al., 2009)	II EP CAPE-V (present study)
Sentence C	**Target:** "All voiced". "We were away a year ago"	**Target:** "Voiced segments". "Olha lá o avião azul" [ˈɔʌɐˈlauɐviɐ̃wɐˈzuɬ] It has only voiced phonemes. However this does not include the voiced EP phonemes [b, d, g, ʒ, m, n, ɲ , l, ʁ].	**Target:** "Only voiced phonemes". "A asa do avião andava avariada" [ɐˈazɐɗuɐviˈɐ̃wɐ̃ɐ̃davɐvɐriˈaɗɐ] It has only voiced phonemes. However it not include the voiced EP [b, g, ʒ, ɲ, m l].	**Target:** "Features all voiced phonemes and provides a context to judge possible voiced stoppages/spasms and one's ability to link from one word to another". "We were away a year ago"	**Target:** Produce all the EP voiced phonemes which allow to judge the possible voiced stoppages/spasms. "A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa" [ɐˈzɛ/ˈmɐ̃j̃dugɐbriˈeɬ/ˈdewʌũˈbo ludɬɨleˈrɐ̃ʒɐiˈviɲuˈvɐʌudɬˈʁunɐ] Sentence included all the EP voiced phonemes.
Sentence D	**Target:** "Elicit hard vocal attacks". "We eat eggs every Easter"	**Target:** "Elicit hard vocal attacks". "Agora é hora de acabar" [ɐˈɣɔɾɐˈɛˈɔɾɐɗɐkɐˈβa] It does not include the hard attack EP vowels [a, i, u].	**Target:** "Hard glottal attack". "Agora é hora de acabar" [ɐˈɣɔɾɐˈɛˈɔɾɐɗɐkɐˈβar] It does not include the hard attack EP [a, i, u].	**Target:** "Includes several vowel-initiated words that may provoke hard glottal attacks and provides the opportunity to assess whether these occur". "We eat eggs every Easter"	**Target:** Sentence that only has words beginning with vowels that elicit glottal attack. "É hora da Urraca ir à caça" [ˈɛˈɔɾɐɗɐuˈʁakɐˈiraˈkasɐ] Sentence that includes all the EP vowels produced in hard glottal attack.

Table 15 (Cont.) – Critical analysis of CAPE-V versions and proposal of new sentences.

	CAPE-V				
	AE CAPE-V – 1st Ed. (ASHA, 2006)	BP CAPE-V (Behlau, 2004)	EP CAPE-V (Jesus et al., 2009a)	AE CAPE-V – 2nd Ed. (Kempster et al., 2009)	II EP CAPE-V (present study)
Sentence E	**Target:** "Incorporates nasal sounds". "My mama makes lemon jam"	**Target:** "Assess nasal sounds emission". "Minha mãe namorou um anjo" [ˈmiɲɐˈmɐ̃ɲɐmuˈroũɐ̃ʒu] It does not include the nasal EP vowels [ĩ, õ, ɐ̃].	**Target:** "Nasal phonemes". "A minha mãe mandou-me embora" [ɐˈmiɲɐˈmɐ̃jmɐ̃domɐ̃bɔrɐ] It does not include the nasal EP vowels [ĩ, õ, ũ] and the consonant [ɲ].	**Target:** "Includes numerous nasal consonants, thus providing an opportunity to assess hyponasality and possible stimulability for resonant voice therapy". "My mama makes lemon jam"	**Target:** Include all the EP nasal vowels and consonants providing the opportunity to assess hyponasality and possible stimulability for resonant voice therapy. "Onde eu brinco há um ninho de andorinhas encostado ao muro" [ˈõˈdewˈbr̃ikuˈaũniɲudɐ̃duˈriɲɐ zɐ̃kuʃˈtaɗwawˈmuru] Sentence includes all the EP nasal vowels and consonants.
Sentence F	**Target:** "Weighed with voiceless plosive sounds". "Peter will keep at the peak"	**Target:** "With voiceless plosive sounds". "Papai trouxe pipoca quente" [pɐˈpajˈtrosɨpiˈpɔkɐˈkẽtɨ] It contains the nasal phoneme [ɐ̃], which is not a target for voiceless plosives.	**Target:** "Voiceless stops". "O Tiago comeu quatro pêras" [uˈtiaɣukuˈmewˈkwatruˈpe reʃ] It contains a nasal phoneme [m] and a voiced plovise sound [g], which is not a target for voiceless plosives.	**Target:** "Contains no nasal consonants and provides a useful context for assessing intraoral pressure and possible hypernasality or nasal air emission". "Peter will keep at the peak"	**Target:** Sentence weighted with voiceless plosive sounds and without any nasal sound to provide a useful context for assessing intraoral pressure and possible hypernasality or nasal air emission. . "A Kika tapou a tua capa preta" [ɐˈkikɐtɐˈpoɐˈtuɐˈkapɐˈpretɐ] Sentence contains the EP voiceless plosive [p, t, k], with an occurrence of three times each.

3.5.2. Voice recording

All the speakers signed the informed consent approved by the Ethics Committee at Hospital da Luz (Appendix A).

All the phonatory tasks were recorded following the CAPE-V instructions (Kempster et al., 2009). A protocol for voice sample collection was developed. This protocol described which and how each voice sample should be collected (Appendix E). Voice samples were recorded in a sound treated room at the ENT Department at Hospital da Luz, with the speakers seated in a comfortable position. The ambient noise was always below 50 dB (Dejonckere et al., 2001), as measured with a digital sound level meter (model Rolls SLM305). Voice productions were recorded directly on the digital recorder TASCAM DR-05, 16 bits, mono, with a sample frequency of 44100 Hz. A PEYLE PMENI headset microphone was positioned at a constant distance of 4 cm from the speaker's mouth and at a 45º angle from the mouth (Dejonckere et al., 2001).

Speakers were asked to sustain [a] and [i] at a steady and comfortable pitch level three times, for 3-5 seconds each time. They were instructed to read aloud the proposed new sentences and to respond to the prompt "Tell me about the place where you grew up" to elicit 20 seconds of spontaneous speech.

The same recording and tasks procedures were used to obtain all the voice samples from all the subjects. Similar to the Zraick et al. (2011) study, voice samples were not normalized for intensity and noise reduction. After each voice sample was recorded, the samples for each subject were labeled with no speaker identification information.

3.5.3. Listening

The 26 voice samples included 10 normal and 10 dysphonic voices and 6 repeated voices (3 normal and 3 dysphonic) randomly mixed to enable test-retest for determining the intra-rater reliability (Zraick et al., 2011). Repeated voice samples were presented to listeners together with the 20 voice samples. Before the first listening session, all the listeners underwent a pure tone hearing screening at -20 dB HL at 500, 1000, 2000, and 4000 Hz (ASHA, 1997). All 26 voice samples were stored in the same pre-established, random sequence (Appendix B). During the first listening session, 14 judges rated the 26

voice samples using the II EP CAPE-V (Appendix D). One week later, they rated the same voice sequence using the GRBAS scale (Annex A) (Mozzanica et al., 2013; Nemr et al., 2012). Voice samples were presented in a quiet room with ambient noise <50 dB, at the ENT Department at Hospital da Luz. Each listener was seated at a computer, equipped with headphones (AKG K101) (Kelchener et al., 2010; Nemr et al., 2012; Patel & Shrivastav, 2007) and was allowed to adjust the volume to a comfortable listening level (Zraick et al., 2011). Each listener was allowed to listen the voice samples more than once (Nemr et al., 2012; Zraick et al., 2011) but no more than 3 times. The voices were reproduced in four blocks of: 1st) seven voice samples, 2nd) six voice samples, 3rd) seven voice samples, 4th) six voice samples, with a 10 minutes interval between each block (Nemr et al., 2012), to reduce fatigue and inattentiveness.

Before the 1st listening session, all the listeners received a complete application manual of the II EP CAPE-V instrument (Appendix F) in order to promote reliability of the voice ratings performed by each listener. The manual contained all information about the parameters and concepts of the instrument, instructions for listening and rating procedures, as well as II EP CAPE-V forms (Appendix D). The listeners were asked to make their judgements based on all the phonatory tasks. Each II EP CAPE-V form was identified with the code number of the voice sample. After listening to each voice sample, each listener marked the deviant degree on the 0-100 millimeter line for each vocal parameter. The listener indicated if the parameter was consistent or intermittent. Resonance was also assessed. Listeners were also encouraged to add two more VQ parameters which they found relevant for that voice sample.

One week later, the same voice sample sequence was rated using the GRBAS scale. Before the voice listening started, the listeners received a complete application manual of GRBAS scale (Appendix G). This included information about vocal parameters and rating procedure, as well as GRBAS forms (Annex A). Each GRBAS form was identified with the code number of the voice sample. The voice samples were reproduced following the same procedures applied in the II EP CAPE-V. After listening the same phonatory tasks of each subject, the listener rated the GRBAS vocal parameters using a Likert scale of 4 points: “0” normal, “1” slight, “2” moderate, and “3” extreme.

Listeners were aware that normal voice samples were included in the random sampling sequence. However, no voice disorder diagnoses were provided to avoid any

bias effect (Eadie et al., 2011a). Listeners were allowed to consult II EP CAPE-V and GRBAS written protocols and definitions at any time, to assist their internal standards.

3.6. Statistical Analysis

Statistical analysis was performed using two statistical software packages: Statistical Package for the Social Sciences 22.0 (IBM SPSS, 2013) and LISREL 8.80 (Jöreskog & Sörbom, 2006).

Construct validity of II EP CAPE-V was based on a contrasted groups approach. The independent-samples Student *t*-test was used to compare means between the CG and the DG, across all the vocal parameters (dependent variables) with α=.05. This analysis was performed using SPSS 22.0 (IBM SPSS, 2013).

The degree of association between the CAPE-V and the GRBAS parameters (concurrent validity) was estimated with a multi-serial correlation coefficient for each listener and for the average scores of the total listeners, with r>.70. This correlation estimates the degree of association between an interval variable (CAPE-V parameters) and an ordinal variable (GRBAS parameters) (Harshbarger, 1977). For this analysis the LISREL software was used.

Inter-rater reliability of the II EP CAPE-V was examined using the ICC calculated following a two-way mixed effects model (Shrout & Fleiss, 1979), with a confidence interval of 95%. Intra-rater reliability was performed with Pearson correlation coefficients (r>.70) for all vocal parameters. For the reliability analyses, all the calculations were performed on SPSS 22.0 statistical software (IBM SPSS, 2013).

IV. RESULTS

The present study was transversal, observational, comparative, and descriptive in nature. 14 SLPs voice experts with ≥5 years of clinical practice rated 20 voice samples produced by 10 males (mean age=45) and 10 females (mean age=43) who were classified into two groups matched by age and gender: CG (n=10) and DG (n=10).

For construct validity analysis, CG and DG mean scores and, standard deviations were compared using independent-sample Student's *t*-test, for all the II EP CAPE-V parameters (Table 16). For all vocal parameters, mean scores and standard deviations of DG were higher than CG. There were significant differences found between DG and CG ($p<.05$) for overall severity, roughness, breathiness, loudness, and pitch. No significant difference between groups was found on the strain parameter. However, it had a higher mean score on DG, and its standard deviation was higher in the CG.

Table 16 – Means and standard deviations of II EP CAPE-V parameters.

	Control group	Dysphonic group	
Vocal parameter	Mean±SD	Mean±SD	*p*-value
Overall severity	12.77 ± 11.88	38.24 ± 21.04	.01*
Roughness	13.68 ± 7.92	39.01 ± 11.49	.00*
Breathiness	12.77 ± 11.88	38.24 ± 21.04	.01*
Strain	23.04 ± 12.87	26.59 ± 11.06	.52
Pitch	7.98 ± 5.18	20.29 ± 10.41	.01*
Loudness	9.62 ± 5.59	20.26 ± 13.59	.04*

SD=standard deviation; $p<.05$.

A multi-serial correlation between the four comparable II EP CAPE-V and GRBAS parameters was determined for each listener, as well as for the average scores of the total of listeners. Overall severity, roughness, and breathiness had the higher correlations ($r>.70$), while strain did not meet this threshold ($r<.50$) (see Table 17).

Table 17 – Multi-serial correlation between II EP CAPE-V and GRBAS parameters.

CAPE-V	GRBAS	Multiserial correlation (range)
Overall severity	Grade	.95 (.22 – .99)
Roughness	Roughness	.89 (.23 – .91)
Breathiness	Breathiness	.90 (.39 – .91)
Strain	Strain	.47 (.18 – .93)

Inter-rater reliability was obtained using ICC for each II EP CAPE-V vocal parameter. There was a high level of agreement (ICC>.84) across all 14 listeners for all the vocal parameters (Table 18). Overall severity presented the highest ICC (ICC=.96) and strain the lowest (ICC=.84).

Table 18 – Inter-rater reliability of II EP CAPE-V parameters.

Vocal parameter	ICC
Overall severity	.96
Roughness	.92
Breathiness	.95
Strain	.84
Pitch	.86
Loudness	.90

ICC=intraclass correlation coefficient.

Six repeated voice samples were used to determine intra-rater reliability of each vocal parameter. Average, highest and lowest individual of intra-rater reliability coefficients (Pearson's *r*) were calculated. Overall severity, breathiness, and pitch parameters revealed high intra-rater reliability (*r*>.84), while strain (*r*=.73) was considered good, and roughness and loudness reflected only moderate intra-rater reliability (*r*=.61, *r*=.69, respectively). Assessing the number of listeners whose intra-rater reliability was higher than .70 is another way to evaluate intra-rater reliability (Table 19). Intra-rater reliability higher than .70 was achieved by at least seven of the fourteen raters on overall severity, breathiness, and loudness are shown in Table 19.

Table 19 – Intra-rater reliability of II EP CAPE-V parameter for the 14 listeners.

Vocal parameters	*r* (range)	No. of rater with *r*>.70
Overall severity	.87 (.38 – .95)	10
Roughness	.61 (.06 – .90)	6
Breathiness	.87 (.01 – .93)	8
Strain	.73 (.21 – .84)	5
Pitch	.92 (.20 – 1.00)	6
Loudness	.69 (.01 – 1.00)	7

In summary, overall severity and breathiness supported concurrent and construct validity, and revealed high inter- and intra-rater reliability. Pitch and loudness also supported concurrent validity, and inter- and intra-rater reliability. Concurrent and construct validity, as well as inter-rater reliability was observed in roughness. The strain

parameter did not support construct or concurrent validity, but it revealed high inter- and intra-rater reliability (see Table 20).

Table 20 – II EP CAPE-V validity and reliability results.

		Validity		Reliability	
		Construct	Concurrent	Inter-rater	Intra-rater
Vocal parameters	Overall severity	✓	✓	✓	✓
	Roughness	✓	✓	✓	X
	Breathiness	✓	✓	✓	✓
	Strain	X	X	✓	✓
	Pitch	✓	NA	✓	✓
	Loudness	✓	NA	✓	✓

✓ – Higher than .70; X – Lower than .70; NA – Not applicable.

V. DISCUSSION

Auditory-perceptual evaluation plays an important role in multidimensional voice evaluation (Carding et al., 2009) and in establishment of a voice therapy plan (Berhman, 2005; Carding et al. 2000). Different scales and schemes are available, that can be selected depending on the clinical or research purposes of the examiners. As recommended by SACMOT (Aaronson et al., 2002), any health status and quality-of-life assessment instrument must be valid and reliable. The CAPE-V (ASHA, 2006) is a more recent auditory-perceptual evaluation instrument compared to the well-known GRBAS scale (Hirano, 1981). Additionally, the CAPE-V has been increasingly used in both clinical and research settings. The CAPE's psychometric characteristics have been reported in several studies to date (Jesus et al, 2009b; Jesus et al., 2009a; Karnell et al., 2007; Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Nemr et al., 2015; Núñez-Batalla et al., 2015; Zraick et al., 2011).

The CAPE-V has been translated into different languages such as BP (Behlau, 2004), EP (Jesus et al., 2009a), IT (Mozzanica et al., 2013) and SP (Núñez-Batalla et al., 2015); its content validity has been supported by different professionals (e.g. SPLs; linguistics; phoniatrics). On the first CAPE-V translation into EP (Jesus et al., 2009a), content validity was indicated by one speech and hearing scientist, one linguist, and three experienced SLPs. Nevertheless, sentence targets established in the original CAPE-V (ASHA, 2006) were not achieved in the first EP translation. In sentence A, all EP nasal vowels were omitted. In sentence B there was one word that did not begin with easy onset. Sentence C missed some of EP voiced phonemes. Sentence D did not included all EP hard glottal attack vowels. Sentence E did not included all EP nasal vowels and consonants. Lastly, sentence F contained nasal and voiced plosive phonemes, which altered the intended phonetic context. Those missed phonemes led to the nonfulfillment of sentence targets. Therefore, this raised content validity problems. In the present study, II EP CAPE-V content validity was assured by an EP linguist expert, who reviewed six new sentences proposed for reading aloud in the sentence task (see Table 11). The sentences proposed for this CAPE-V version were designed to accomplish the same purposes and phonetic environments stated in the AE 2nd edition of CAPE-V (Kempster et al., 2009). Sentence A assured coarticulatory production of all nasal and oral EP vowels. Sentence B contained only words that begin with easy onset to assess soft glottal

attacks. Sentence C included all EP voiced phonemes to assess possible voiced stoppages/spams. Sentence D included all EP vowels that elicit hard glottal attack. Sentence E contained all EP nasal vowels and consonants to assess nasality. Sentence F included many voiceless plosives to assess intraoral pressure. This guaranteed that the six new sentences fulfilled the target objectives established under AE 2nd edition of CAPE-V (Kempster et al., 2009), as assessed by EP a linguist within a cultural context. For spontaneous speech elicitation, the prompt "Tell me about the place where you grew up" was used, similar to the standardized procedures of CAPE-V (Zraick et al., 2011).

Content validity evidence assured the Portuguese clinicians that the II EP CAPE-V reading aloud task measures the same sentence targets and phonemic environments established under the AE 2nd edition of CAPE-V (Kempster et al., 2009), and that the spontaneous speech task is elicited with the same procedure, regardless of the language in which CAPE-V has been translated. This psychometric characteristic allows for a valid comparison of clinical and research results in the assessment of VQ as reported in different national and international studies.

To establish construct validity of II EP CAPE-V, the mean scores from the CG and the DG were compared for all the CAPE-V parameters. Statistically significant differences ($p<.05$) were found between the two groups for overall severity, roughness, breathiness, pitch, and loudness parameters. This was similar to the results reported by Mozzanica et al. (2013) and Nerm et al. (2015). The strain parameter also revealed differences between the CG and DG. Strain mean score was higher for the DG (mean=26.59) than for the CG (mean=23.04) as expected, which possibly contributes to the identification of a voice disorder for a given speaker. This parameter is usually rated based on a listener's auditory perception added to visual perception of neck muscle tension. In the current study, only auditory stimuli were provided. This result suggests that strain is not a valuable auditory-perceptual parameter to differentiate normal or dysphonic VQ of EP population. Surprisingly, from all the voice parameters, strain had the highest mean scores (ranged from 7.98 to 23.04) and standard deviations (ranged from 5.18 to 12.87) in the CG, similar to those reported by Nerm et al. (2015). This result could be influenced by the fact that listeners were aware that voice samples included normal and dysphonic voices. Nonetheless, no voice disorder diagnoses were provided in order to avoid a bias effect in judging (Eadie et al., 2011a). Another possible reason for these

results could be that CAPE-V used a VAS for vocal parameter ratings, which allows for a more detailed analysis compared to an ordinal scale.

Results reported in this study support that II EP CAPE-V is a valid instrument for identification and characterization of normal and dysphonic speakers, and is able to distinguish them except on the parameter of strain. This study is innovative and relevant for both national and international clinical and research endeavors because it contributed to establishing CAPE-V construct validity, where little data are available.

In future research, it would be helpful to study the sensitivity of the CAPE-V in order to document VQ improvement during voice therapy. For that purpose, voice samples recorded at the beginning and end of voice therapy should be rated by different SLPs who are experts in voice disorders, using the CAPE-V. Further studies could include auditory and visual stimuli together, in order to a better understanding of the dimension of strain and how it is evaluated. It would be also interesting to study the correlation between electromyography findings and the auditory-perceptual evaluation of strain in normal voices. Using the CAPE-V to evaluate different laryngeal disorders (e.g. structural pathologies; inflammatory condition; neurological disorders) could be helpful in order to observe what CAPE-V parameters better characterize and distinguish those disorders. The mean scores obtained in the present study for all the vocal parameters in the CG may be interpreted as supporting the need for EP SLPs' training in auditory-perceptual parameters presents in normal and dysphonic voices. A training course about auditory-perceptual evaluation should be included in EP SLP graduate programs, and EP SLP experts in voice disorders should also take a training course to refresh their internal standards.

Concurrent validity was established in this investigation based on the multi-serial correlations between the four comparable II CAPE-V and GRBAS parameters: overall severity/grade, roughness, breathiness, and strain. Listeners rated the voices first using the II EP CAPE-V and one week later using the GRBAS, avoiding a potential cross-over effect. Results revealed high correlations between overall severity/grade (r=.95), roughness (r=.89), and breathiness (r=.90). These results were similar to those reported by Karnell et al. (2007), and higher than the reported by Jesus et al. (2009b), Mozzanica et al. (2013), Núñez-Batalla et al. (2015), and Zraick et al. (2011) (see Table 21). The II EP CAPE-V and GRBAS strain correlation was lower (r=.47) than that reported in the

Karnell et al. (2007), Mozzanica et al. (2013), Núñez-Batalla et al. (2015), and Zraick et al. (2011) studies. This result seems to be in agreement to what was found for construct validity. Strain was an auditory-perceptual parameter difficult to measure by EP listeners, with no significant difference found between the CG and DG. This finding may also be influenced by the type of rating scale (VAS vs ordinal scale) used by these instruments. When a larger number of levels of ratings are available, a lack of consistency with some random errors are observed (Kreiman et al., 1993; Wuyts et al., 1999).

The results reported here support the evidence that CAPE-V and GRBAS measure similar constructs, contributing to the establishment of II EP CAPE-V concurrent validity. These results have an impact on clinical practice because they support the use of CAPE-V for auditory-perceptual voice evaluation and voice therapy outcomes measurement in national and international studies. The CAPE-V has formal administration procedures with determined phonatory tasks, encouraging clinicians to follow a standard auditory-perceptual voice evaluation protocol. This instrument uses a VAS to rate more vocal parameters than the GRBAS, which allows for a more detailed VQ evaluation. When selecting either of these two instruments, the user must consider the psychometric characteristics as well as the advantages and disadvantages of both, depending on the purpose of the assessment.

Further investigation is needed to understand if the low strain correlation results from: 1) type of scale (VAS vs ordinal scale); or 2) other inherent difficulties with rating this parameter. It would be helpful to study the strain parameter results when different phonatory tasks are rated with the CAPE-V and GRBAS, in EP speakers' voices. Applying the CAPE-V or GRBAS to evaluate the three phonatory tasks separately may help to find if strain ratings are similar across the phonatory tasks. The present results also support the need for auditory-perceptual evaluation training, especially with respect to strain.

Table 21 – CAPE-V concurrent validity measured with CAPE-V and GRBAS instruments.

		Jesus et al. (2009b)	Karnell et al. (2007)	Zraick et al. (2011)	Nerm et al. (2012)	Mozzanica et al. (2013)	Núñez-Batalla et al. (2015)	Present study
	Statistics	ρ	*r*	*r*	*r*	*r*	ICC	*r*
Vocal parameters	Overall severity/grade	ρ=.60	*r*=95	*r*=.80	*r*=.80	*r*=.92	ICC=.874	*r*=.95
	Roughness	ρ=.26	*r*=.90	*r*=.76	NA	*r*=.84	ICC=.849	*r*=.89
	Breathiness	ρ=.80	*r*=.89	*r*=.78	NA	*r*=.87	ICC=.612	*r*=.90
	Strain	NA	*r*=.91	*r*=.77	NA	*r*=.79	ICC=.843	*r*=.47

NA – Not available.

Reliability is a necessary psychometric measure of the validity of an instrument because it establish the degree in which an instrument is free from random error, and the extent to which results can be reproduced. In the current study, inter- and intra-rater reliability were analyzed across 14 listeners for all vocal parameters.

High inter-rater reliability (ICC>.84) was found for all parameters (see Table 22), demonstrating strong agreement among 14 listeners. Compared to these results, Jesus et al. (2009a) reported similar inter-rater reliability results for overall severity, and higher results for the breathiness and loudness parameters. However, in their study only two listeners rated 10 disordered voice samples, and inter-rater reliability was calculated using Spearman's correlation coefficient. In the present study, ICC was calculated to determine inter-rater reliability across a larger number of listeners (14 listeners) who rated 20 voice samples (10 normal and 10 dysphonic). Apart from the Jesus et al. (2009a) study, inter-rater reliability reported in this study revealed the highest correlation of agreement for all the vocal parameters, when compared to what has been reported in other studies (see Table 16). These results may be due to the larger number of listeners used. Most of the CAPE-V studies had a maximum of 4 listeners, while in this study there were 14. The Zraick et al. (2011) was the study with the larger number of listeners (21). Nevertheless, the number of EP voice experts in Portugal is lower than in USA, thus this factor does not diminish inter-reliability value. Inter-rater reliability can also be influenced by a listener's experience. In the current study, the 14 listeners were SLPs and experts in voice disorders, with more than five years of clinical practice. Listeners' experiences and clinical backgrounds do influence inter-rater reliability (Bassich & Ludlow, 1986; Eadie et al., 2010; Helou et al., 2010; Kreiman et al., 1990; Kreiman et al., 1993; Kreiman et al., 1992; Sofranko & Prosek, 2012). Experienced listeners usually reveal better inter-rater reliability compared to inexperienced listeners (De Bodt et al., 1997; Helou et al., 2010; Sofranko & Prosek, 2012; Zraick et al., 2005). However, in the present study the listeners were selected with the understanding that these factors influence inter-rater reliability results. Furthermore, in the current study, overall severity, roughness, and breathiness were the vocal parameters with highest inter-rater reliability, similar to those reported by Kelchener et al. (2010), Mozzanica et al. (2013), Nerm et al. (2012), Núñez-Batalla et al. (2015), and Zraick et al. (2011),. These results are in agreement with evidence that expert listeners' inter-rater reliability is higher for the overall severity, roughness, and breathiness parameters (De Bodt et al., 1997; Chan & Yiu, 2006; Iwarson

& Peterson, 2012; Karnell et al., 2007; Kreiman & Gerratt, 1998; Webb, Carding et al., 2004).

Voice stimuli may also influence the high inter-rater reliability achieved in this study. Voice stimuli included the three CAPE-V phonatory tasks, produced by 10 normal and 10 dysphonic subjects, matched for age and gender. These balanced voice stimuli may have contributed to low variability across the CG and DG, resulting in a better inter-rater agreement. Results reported in this study support the II EP CAPE-V inter-rater reliability. This psychometric characteristic is particularly important because it demonstrated that 14 EP expert listeners rated CAPE-V vocal parameters consistently, independently of listeners' different backgrounds, clinical settings, and internal standards. This indicates that II EP CAPE-V results are similar to those reported in other international CAPE-V studies, which allows for the sharing and comparison of auditory-perceptual results from various national or international studies.

Further investigation using inexperienced listeners is needed to better understand the impact a listener's experience has in the II EP CAPE-V auditory-perceptual voice evaluation. It would also be helpful to establish the number of listeners that may best allow for adequate reliability of auditory-perceptual parameter evaluation. The impact of the different phonatory tasks on the CAPE-V inter-reliability results is also worthy of investigation. CAPE-V phonatory tasks could be rated all together at the same time, and separated apart with one week interval between each one, in order to determine the phonatory tasks' impact in auditory-perceptual reliability results.

The II EP CAPE-V revealed high intra-rater reliability for overall severity (r=.87), breathiness (r=.87), and pitch (r=.92); good reliability (r=.73) for strain; and moderate reliability for roughness (r=.61) and loudness (r=.69). These findings revealed the stability of each listener's rating for those vocal parameters (see Table 23). In general, these results were lower than those reported by Mozzanica et al. (2013) and Núñez-Batalla et al. (2015), with exception of the strain parameter, which had higher results here. These differences may result from the methodological procedures applied. In both studies, voice samples were re-rated with a one week interval, and intra-rater reliability was determined based on the ICC results. However, the number of listeners varied: three listeners were used in Mozzanica et al. (2013) study and one single listener in the Núñez-Batalla et al. (2015) study. Intra-rater reliability can be influenced by a listener's internal

standards (Kreiman, Gerratt & Ito, 2007; Kreiman et al., 2004), which change accordingly to the listener's previous voice experience, as well as the listener's memory (Gerratt et al., 1993; Kreiman et al., 1993; Kreiman et al., 1992; McAlliser, Sundberg & Hibi, 1996). When controlled, those factors have less of an influence on intra-rater reliability. In the present study, six repeated voice samples were presented to 14 listeners together with the total of 20 voice samples, all randomly mixed, similar to the methodology adopted by Zraick et al. (2011). Breathiness and loudness revealed similar intra-rater reliability results to the reported by Zraick et al. (2011), with at least half of listeners achieving high reliability in both studies. For current study, overall severity, strain, and pitch revealed higher intra-rater reliability than those reported by Zraick et al. (2011), while roughness was lower. These results could have been influenced by different factors such as the: number of repeated voice samples (6 vs 11 respectively), the number of phonatory tasks rated (spontaneous speech vs three CAPE-V phonatory tasks), or the rating session methodology adopted. In the present study, 30% of total voice samples were re-rated by listeners, while in Zraick et al. (2011) study 18% were re-rated. This factor may decrease the intra-rater reliability representativeness. Differences related to the phonatory task ratings might also have had an impact on intra-reliability results. Even if spontaneous speech is more reliable than ratings of sustain vowels (Bele, 2005; Eadie & Doyle, 2005; Law et al., 2012; Zraick et al., 2005), a complete voice evaluation should always include both phonatory tasks (Maryn & Roy, 2012). In the present study, all voice samples were rated in two sessions with a one week interval. In first session, all voice samples were rated with II EP CAPE-V, while in second with the GRBAS, guaranteeing that listeners experienced the same conditions, thus, minimizing at possible internal standards changing over time. In contrast, the Zraick et al. study (2011), divided listeners into two groups: Group A, which applied GRBAS scale in the first rating session and CAPE-V in the second one; and Group B, which applied the CAPE-V in the first rating session, and GRBAS in the second one; both sessions separated by 48-72 hours. This methodology did not assure that listener's internal standards remained similar across two rating sessions. In current study, pitch intra-rater reliability (r=.92) was higher than the reported in literature (Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015; Zraick et al., 2011). This result showed that pitch was a remarkable and stable auditory-perceptual parameter for EP listeners. In this study, intra-rater variability found for each listener could be influenced by a listener's experience (Eadie et al., 2010; Helou et al., 2010; Kreiman et al., 1990; Kreiman et al., 1993; Kreiman et

al., 1992; Sofranko & Prosek, 2012). However, the 14 listeners were voice experts with at least five years of experience, similar to most of CAPE-V intra-rater reliability studies (Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Zraick et al., 2011).

This study was the first reporting CAPE-V intra-rater reliability when applied to EP voice samples by EP listeners. The same voice samples sequence was presented to listeners, and intra-rater reliability was measured based on the results obtained for six repeated voice samples (30% of total). Results indicated that EP listeners displayed stable internal standards, demonstrating their intra-rater reliability for auditory-perceptual VQ evaluation. Overall severity, breathiness, strain, and pitch were the II EP CAPE-V parameters with the highest agreement among repeated voice samples, revealing that listeners rated those vocal parameters consistently. The EP II CAPE-V intra-rater reliability reported promotes its use in both clinical and research auditory-perceptual voice evaluation, because it indicated that vocal parameters were constantly rated by experienced listeners, independent of the rating moment.

Further investigation is needed to clarify if intra-rater reliability is influenced by the presence or absence of a voice disorder. It would be also worthwhile to further study a listener's reliability among the three phonatory tasks and observe if there is one task that may display stronger test-retest features than the others.

Content validity was supported in the II EP CAPE-V. The significant differences ($p<.05$) found between the control and dysphonic group for overall severity, roughness, breathiness, pitch, and loudness supported II EP CAPE-V construct validity, accepting the alternative hypothesis for tested hypothesis 1. These results indicated that those were the vocal parameters that better distinguished normal from dysphonic voices. The high correlation coefficients ($r>.70$) between the CAPE-V and GRBAS parameters of overall severity/grade, roughness, and breathiness revealed the II EP CAPE-V concurrent validity. Therefore, the alternative hypothesis for hypothesis 2 was accepted, indicating that both instruments measure similar constructs. High level of agreement (ICC>.70) between the listeners in all vocal parameter supported II CAPE-V inter-rater reliability. The alternative hypothesis for hypothesis 3 was accepted, indicating that listeners were reliable in their voice sample ratings. The high level of agreement ($r>.70$) among the repeated voice samples ratings by each listener demonstrated II CAPE-V intra-rater reliability for overall severity, breathiness, strain, pitch, and loudness parameters.

Therefore, the alternative hypothesis for hypothesis 4 was accepted, indicating the EP SLPs were reliable in their ratings.

Limitations of this study can be related to methodological procedures. A smaller number of voice samples was used in comparison to other CAPE-V studies (Jesus et al., 2009b; Karnell et al., 2007; Kelchener et al., 2010; Mozzanica et al., 2013; Nerm et al., 2012; Núñez-Batalla et al., 2015; Zraick et al., 2011). Nevertheless, the speakers were selected according to the laryngoscopy results and were matched for age and gender. All listeners were experts in voice disorders with an average of 11 years of clinical practice. This does not assure II EP CAPE-V acceptable validity and reliability when used by inexperience listeners. No anchor stimuli were provided before the II EP CAPE-V rating session. Reliability results could have been influenced by this because the CAPE-V is a recent instrument, and the EP listeners were not accustomed to its use. Voice stimuli were comprised of the three phonatory tasks established by the II EP CAPE-V. This may had some impact in the validity and reliability reported, although the procedures used here allowed listeners to perform a global evaluation of each voice sample according to the rationale and closely following the protocol of the CAPE-V authors.

Table 22 – Inter-rater realibility across CAPE-V studies.

		Jesus et al. (2009a)	Karnell et al. (2007)	Kelchner et al. (2010)	Zraick et al. (2011)	Nerm et al. (2012)	Mozzanica et al. (2013)	Núñez-Batalla et al. (2015)	Present study
	Statistics	ρ	*r*	ICC	ICC	ICC	ICC	ICC	ICC
Vocal parameters	Overall severity	ρ=.964	*r*>.88	ICC=67%	ICC=.76	ICC=.911	ICC=.92	ICC>.833	ICC=.96
	Roughness	ρ=.834	NA	ICC=68%	ICC=.62	ICC=.870	ICC=.91	ICC>.750	ICC=.92
	Breathiness	ρ=.991	NA	ICC=71%	ICC=.60	ICC=.897	ICC=.90	ICC>.769	ICC=.95
	Strain	ρ=.659	NA	ICC=35%	ICC=.56	ICC=.828	ICC=.76	ICC>.648	ICC=.84
	Pitch	k=0.500	NA	ICC=68%	ICC=.54	NA	ICC=.83	ICC>.710	ICC=.86
	Loudness	k=1.000	NA	ICC=63%	ICC=.28	NA	ICC=.82	ICC>.545	ICC=.90

NA – Not available.

Table 23 – Intra-rater reliability across CAPE-V studies.

		Karnell et al. (2007)	Kelchner et al. (2010)	Zraick et al. (2011)	Nerm et al. (2012)	Mozzanica et al. (2013)	Núñez-Batalla et al. (2015)	Present study
	Statistics	*r*	ICC	*r*	ICC	ICC	ICC	*r*
Vocal parameters	Overall severity	*r*>.88	ICC=87%	*r*=.57	ICC=.927	ICC=.92	ICC>.972	*r*=.87
	Roughness	NA	ICC=82%	*r*=.77	NA	ICC=.92	ICC>.969	*r*=.61
	Breathiness	NA	ICC=82%	*r*=.82	NA	ICC=.90	ICC>.952	*r*=.87
	Strain	NA	ICC=63%	*r*=.35	NA	ICC=.89	ICC>.921	*r*=.73
	Pitch	NA	ICC=78%	*r*=.78	NA	ICC=.88	ICC>.894	*r*=.92
	Loudness	NA	ICC=79%	*r*=.64	NA	ICC=.80	ICC>.851	*r*=.69

NA – Not available.

CONCLUSION

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

II EP CAPE-V content validity was obtained by reading aloud and spontaneous speech tasks, contributing for the CAPE-V standardization regardless of a translation's language.

II EP CAPE-V construct validity was assured, revealing that overall severity, roughness, breathiness, pitch, and loudness were the vocal parameters that distinguish normal and dysphonic voices.

II EP CAPE-V concurrent validity was supported by the high correlation achieved between the CAPE-V and GRBAS overall severity/grade, roughness, and breathiness. The selection of each instrument should depend on the clinical or research purpose of the auditory-perceptual VQ evaluation.

High inter- and intra-rater reliability reported emphasizes II EP CAPE-V reproducibility. In general, overall severity, breathiness, and pitch had high inter and intra-rater reliability, demonstrating that these are the most valuable auditory-perceptual parameters for VQ evaluation. Roughness, strain, and loudness are more salient for auditory-perceptual evaluation across listeners than within the same solo listener.

ANNEXES

ANNEX A: GRBAS (Hirano, 1981).

Escala GRBAS

G____ R____ B____A____S____

ANNEX B: CAPE-V (ASHA, 2006).

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Name:_____ **Date:**_____

The following parameters of voice quality will be rated upon completion of the following tasks:
1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent I = Intermittent
MI = Mildly Deviant
MO =Moderately Deviant
SE = Severely Deviant

			SCORE
Overall Severity	_____ (MI MO SE)	C I	/100
Roughness	_____ (MI MO SE)	C I	/100
Breathiness	_____ (MI MO SE)	C I	/100
Strain	_____ (MI MO SE)	C I	/100
Pitch	(Indicate the nature of the abnormality): _____ _____ (MI MO SE)	C I	/100
Loudness	(Indicate the nature of the abnormality): _____ _____ (MI MO SE)	C I	/100
_____	_____ (MI MO SE)	C I	/100
_____	_____ (MI MO SE)	C I	/100

COMMENTS ABOUT RESONANCE: NORMAL OTHER (Provide description):_____

_____

ADDITIONAL FEATURES (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, or other relevant terms):

Clinician:_____

ANNEX C: 1st EP version of CAPE-V (Jesus et al, 2009a).

Os parâmetros da qualidade vocal que se seguem, devem ser avaliados com recurso às seguintes tarefas:

1. Vogais sustentadas /a/ e /i/ (três repetiçoões com duração de 3-5 segundos cada)
2. Leitura de frases:
   a. A Marta e o avô vivem naquele casarão rosa velho;
   b. Sofia saiu cedo da sala;
   c. A asa do avião andava avariada;
   d. Agora é hora de acabar;
   e. A minha mãe mandou-me embora;
   f. O Tiago comeu quatro pêras;
3. Fala espontânea (minimos 20 seg)
   a. Novo paciente: "fale-me como começou o seu problema de voz, quando notou e o que fez em relação a isso";
   b. Paciente em acompanhamento: "diga-me como está a sua voz" (com duração de 20 segundos)

Legenda: C = consistente I = inconsistente
AL: alteração ligeira
AM: alteração moderada
AS: alteração severa

Pontuação

Grau de severidade Global _____ C I ____/100
AL AM AS

Rouquidão _____ C I ____/100
AL AM AS

Soprosidade _____ C I ____/100
AL AM AS

Tensão _____ C I ____/100
AL AM AS

Altura tonal (indicar o tipo de alteração) _____

_____ C I ____/100
AL AM AS

Intensidade (indicar o tipo de alteração) _____

_____ C I ____/100
AL AM AS

RESSONÂNCIA: Normal Alterada (breve descrição): _____
_____

OUTROS PARÂMETROS (por ex.: diplofonia, aspreza, falseto, astenia, afonia, bitonalidade, tremor, estridente, "glottal fry", entre outros aspecotos relevantes) _____
_____

ANNEX D: ASHA permission to translate the CAPE-V into EP for use in this study.

AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION

June 16, 2015

To Whom It May Concern:

Non-exclusive permission is granted for Sancha C. de Almeida to translate and reprint the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) instrument, descriptions, and instructions into European Portuguese for use in the research project, "Validity and Reliability for the 2nd EP Version of the CAPE-V."

Sincerely,

Libby Bauer

Libby Bauer
Subscription & Permissions Manager
permissions@asha.org

2200 RESEARCH BOULEVARD ROCKVILLE, MARYLAND 20850-3289 301-296-5700 VOICE 301-296-5777 FAX www.asha.org

APPENDICES

APPENDIX A: Speakers informed consent form

Formulário de Consentimento

Investigação: Validade e fidelidade da 2ª versão do instrumento de avaliação "*Consensus Auditory-Perceptual Evaluation of Voice*" para o Português Europeu (II CAPE-V PE).

Equipa de Investigação: Sancha Almeida (916 309 013/scalmeida@hospitaldaluz.pt); Ana Brito Mendes

Agradecemos por participar voluntariamente neste projecto de investigação. O objectivo deste formulário é explicar por escrito em que consiste este projecto de investigação para que possa de modo informado, dar o seu consentimento, assinando o presente documento.

Este projecto tem como principal objectivo contribuir para a validação da 2ª versão do instrumento de avaliação áudio-perceptiva da voz "*Consensus Auditory-Perceptual Evaluation of Voice*" para o Português Europeu (PE) (II CAPE-V PE). Através deste estudo, pretende-se promover a uniformização da avaliação audio-perceptiva da voz de todos os utentes por parte dos clínicos especialistas. Este estudo é um estudo transversal, descritivo, observacional e comparativo.

A sua voz será gravada durante a produção de vogais, leitura de frases e discurso espontâneo. Posteriormente, 14 terapeutas da fala especialistas em voz irão proceder à análise áudio-perceptiva da sua voz nas diferentes tarefas gravadas.

As tarefas de voz referidas serão gravadas utilizando um microfone de cabeça. Estes procedimentos **não são invasivos e não têm quaisquer riscos associados**. A gravação demora cerca de 10 minutos e será feita sentada.

O tempo médio previsto de **recolha de dados** será de aproximadamente **15 minutos**. Estes registos serão arquivados no Hospital da Luz, estando a sua consulta **reservada apenas aos membros da equipa de investigação do projecto**. Quaisquer dados pessoais são **confidenciais**, pelo que não serão divulgados em apresentações ou

publicações resultantes deste projecto. Na condução da investigação, a **total segurança dos sujeitos é salvaguardada** durante todo o processo.

Finalmente, gostaria de o(a) informar que, a qualquer momento, **pode desistir** da sua participação nesta investigação **sem qualquer penalização ou obrigação** para com a equipa de investigação. Se tiver perguntas, comentários ou recomendações sobre a mesma pode contactar o investigador.

Eu, **(letras maiúsculas e de imprensa)**_____ _____, declaro que li e compreendi a informação acima descrita e, voluntariamente, participo neste projecto. Compreendo que **não há remuneração ou compensações por esta participação**. Compreendo também, que os registos são totalmente confidenciais e tenho o direito de desistir desta participação a qualquer momento.

Recebi e assinei este formulário por concordar com as condições deste projecto.

_____

(assinatura do sujeito participante)

Número de identificação atribuído ao sujeito: _____

Certifico que expliquei a natureza e o objectivo deste estudo, os potenciais benefícios e riscos associados à participação neste projecto de investigação. Respondi a todas as questões colocadas pelo sujeito participante.

_____, ___ de _____ de _____

_____

(assinatura de um membro da equipa de investigação)

APPENDIX B: Voice stimuli characterization.

Table B.1 - Voice stimuli characterization.

Voice sample number	Age	Gender	ENT Diagnosis	Classification Manual for Voice Disorders – I [(1)]
1	34	Male	Normal exam	
2	42	Male	Left vocal fold paresis	Neurologic disorder
3	42	Male	Normal exam	
4	44	Female	Right vocal fold paresis	Neurologic disorder
5	61	Male	Laryngopharyngeal reflux	Inflammatory disorder
6	37	Male	Normal exam	
7	30	Female	Vocal fold nodules	Structural pathologies
8 (Repetition of sample 2)	---------	---------	---------	---------
9	34	Female	Normal exam	
10 (Repetition of sample 7)	---------	---------	---------	---------
11	37	Male	Left vocal fold nodule	Structural pathologies
12	44	Female	Normal exam	
13 (Repetition of sample 1)	---------	---------	---------	---------
14	61	Male	Normal exam	
15	52	Female	Normal exam	
16	52	Male	Normal exam	
17	34	Male	Bilateral vocal fold sulcus	Structural pathologies
18	34	Female	Vocal fold nodules	Structural pathologies
19 (Repetition of sample 15)	---------	---------	---------	---------
20	55	Female	Normal exam	
21	55	Female	Arytenoid asymmetric movement with glottis chick	Other disorder
22	30	Female	Normal exam	
23 (Repetition of sample 18)	---------	---------	---------	---------
24	52	Male	Ventricular dysphonia	Other disorder
25 (Repetition of sample 3)	---------	---------	---------	---------
26	52	Female	Reinke's edema	Structural pathologies

[(1)]Verdolini, Rosen & Branski (2006); ENT – ear, nose and throat.

APPENDIX C: Listeners informed consent form

Formulário de Consentimento

Investigação: Validade e fiabilidade da 2ª versão do instrumento de avaliação "*Consensus Auditory-Perceptual Evaluation of Voice*" para o Português Europeu (II CAPE-V PE).

Equipa de Investigação: Sancha Almeida (916 309 013/scalmeida@hospitaldaluz.pt); Ana Brito Mendes

Agradecemos por participar voluntariamente neste projecto de investigação. O objectivo deste formulário é explicar por escrito em que consiste este projecto de investigação para que possa de modo informado, dar o seu consentimento, assinando o presente documento.

Este projecto tem como principal objectivo contribuir para a validação da 2ª versão do instrumento de avaliação áudio-perceptiva da voz "*Consensus Auditory-Perceptual Evaluation of Voice"* para o Português Europeu (PE) (II CAPE-V PE). Através deste estudo, pretende-se promover a uniformização da avaliação audio-perceptiva da voz de todos os utentes por parte dos clínicos especialistas. Este estudo é um estudo transversal, descritivo, observacional e comparativo.

Ser-lhe-á fornecido um manual de aplicação do II CAPE-V PE, com informação relativa aos parâmetros e conceitos do instrumento, directrizes relativas à forma classificar as amostras de voz e número suficiente de cópias de folhas de registo do instrumento II CAPE-V PE.

Numa sala silenciosa, serão ouvidas o total de 26 amostras de voz referentes a 20 sujeitos com voz normal e/ou disfónica. Cada amostra de voz, conterá todas as tarefas avaliadas pelo instrumento II CAPE-V PE (produção de vogais, leitura de frases e discurso espontâneo).

As tarefas de voz referentes a cada sujeito serão ouvidas numa sala silenciosa usando uns **auscultadores com um volume confortável** (determinado por cada juiz) e

poderão ser ouvidas até três vezes no máximo. As vozes deverão ser **reproduzidas em quatro blocos** com intervalo de, no mínimo, 10 minutos entre cada bloco. Cada bloco é composto por: 1) 7 amostras de voz, 2) 6 amostras de voz, 3) 7 amostras de voz, e 4) 6 amostras de voz. Após ouvir todas as tarefas que compõem a amostra de voz referente a um sujeito, deverá proceder à avaliação áudio-perceptiva global da voz preenchendo uma folha de registo do instrumento II EP CAPE-V EP. O preenchimento da folha será feito da seguinte forma: cada parâmetro da qualidade vocal (grau de severidade global, rouquidão, soprosidade, tensão, altura tonal e intensidade) deverá ser avaliado através de uma marca vertical ao longo da linha de 100 mm (o = sem alteração; 100 = fortemente alterada). De seguida, terá que indicar a consistência da presença de cada um dos parâmetros avaliados. Deve ainda avaliar a ressonância e poderá ainda adicionar outros parâmetros que julgue serem relevantes na avaliação da amostra de voz. Uma semana depois, deverá proceder-se à avaliação áudio-percetiva das mesmas amostras de voz através da escala GRBAS. Ser-lhe-á fornecido um manual de aplicação da GRBAS, com informação relativa aos parâmetros e conceitos da mesma assim como número suficiente de cópias de folhas de registo. A reprodução das vozes será feita de forma idêntica à reprodução para a avaliação com o instrumento II CAPE-V PE. O preenchimento da folha de registo da GRBAS será feito da seguinte forma: cada parâmetro da qualidade vocal (grau geral da alteração vocal, rouquidão, soprosidade, astenia e tensão) deverá ser avaliado numa escala de likert de 4 pontos sendo "0" normal, "1" alteração ligeira, "2" alteração moderada e "3" alteração severa.

Estes procedimentos **não são invasivos e não têm quaisquer riscos associados**. A audição e avaliação de cada bloco de amostras terá duração de cerca de 20 minutos e será feita sentada.

O tempo médio previsto para a audição e avaliação de todas as amostras de voz **será** de aproximadamente **2 horas** feita em **2 sessões**. Estes registos serão arquivados no Hospital da Luz, estando a sua consulta **reservada apenas aos membros da equipa de investigação do projecto**. Quaisquer dados pessoais são **confidenciais**, pelo que não serão divulgados em apresentações ou publicações resultantes deste projecto. Na condução da investigação, a **total segurança dos sujeitos é salvaguardada** durante todo o processo.

Finalmente gostaria de o(a) informar que, a qualquer momento, **pode desistir** da sua participação nesta investigação **sem qualquer penalização ou obrigação** para com

a equipa de investigação. Se tiver perguntas, comentários ou recomendações sobre a mesma pode contactar o investigador.

Eu, **(letras maiúsculas e de imprensa)**_____ _____, declaro que li e compreendi a informação acima descrita e, voluntariamente, participo neste projecto. Compreendo que **não há remuneração ou compensações por esta participação**. Compreendo também, que os registos são totalmente confidenciais e tenho o direito de desistir desta participação a qualquer momento.

Recebi e assinei este formulário por concordar com as condições deste projecto.

_____

(assinatura do sujeito participante)

Número de identificação atribuído ao sujeito júri: _____

Certifico que expliquei a natureza e o objectivo deste estudo, os potenciais benefícios e riscos associados à participação neste projecto de investigação. Respondi a todas as questões colocadas pelo sujeito participante.

_____, ___ de _____ de _____

_____

(assinatura de um membro da equipa de investigação)

APPENDIX D: II EP CAPE-V form.

Amostra de voz #_____ **Data aplicação ___/___/___**

Os parâmetros da qualidade vocal devem ser medidos recorrendo às seguintes tarefas fonatórias:

1. **Vogais sustentadas /a/ e /i/** (três repetições de 3-5 segundos cada)
2. **Leitura de frases**:
   a. Num domingo esteve sol e fui com o avô António à esplanada "Évora" comer uma empada.
   b. Segundo Simão, só Samuel sabe.
   c. A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa.
   d. É hora da Urraca ir à caça.
   e. Onde eu brinco há um ninho de andorinhas encostado ao muro.
   f. A Kika tapou a tua capa preta.
3. **Discurso espontâneo** (mínimo 20 seg.) "Fale-me do sítio onde cresceu"

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

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

				Pontuação
Grau de severidade global	_____ C	I	____/100	
	DL DM DS			
Rouquidão	_____ C	I	____/100	
	DL DM DS			
Soprosidade	_____ C	I	____/100	
	DL DM DS			
Tensão	_____ C	I	____/100	
	DL DM DS			

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

_____ C	I	____/100	
DL DM DS			

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

_____ C	I	____/100	
DL DM DS			
_____ _____ C	I	____/100	
DL DM DS			
_____ _____ C	I	____/100	
DL DM DS			

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

_____

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

_____

APPENDIX E: Manual of procedures for voice data collection – II EP CAPE-V.

Manual of procedures for voice data collection II EP CAPE-V

The speaker should be seated comfortably in a quiet environment. The clinician audio-records speaker's performance on the following three phonatory tasks: vowels, sentences, and spontaneous speech. Before the voice recording start the clinician should calibrate and verify all the standard recording procedures.

a) <u>Standard recording procedures:</u>

1. Connect the headset microphone PEYLE PMENI to the digital recorder TASCAM DR-05;
2. Turn on the digital recorder TASCAM DR-05;
3. Press the record bottom of the TASCAM DR-05 to verify the following record settings displayed on the screen:
   3.1. sampling rate – 44100 Hz;
   3.2. file format – WAV;
   3.3. type – mono;
   3.4. resolution – 16 bits;

If any of these settings is not defined as it is described above, the setting must be changed. For that, press the bottom "menu" » "rec setting" » select the setting that are needed to be correct.

4. Place the headset microphone PEYLE PMENI and the digital sound level meter ROLLS SLM305 next to each other and in 4 cm distance from a sound column – see image scheme below. Then produce pure tone of 500 Hz for 5 seconds and record it in the TASCAM DR-05.After that, verify if the fundamental frequency of the calibration sound sample is 500 Hz. For that use the PRAAT software.
5. Place the headset microphone PEYLE PMENI 45 degrees off from of the mouth and 4 cm from the speaker's mouth;
6. Measure the ambient noise with the digital sound level meter ROLLS SLM305. It should be lower then 50 dB;

b) <u>Voice recording</u>

1. Task 1: <u>Sustained vowels</u>

The clinician should say to the speaker "The first task is to say the sound /a/. Hold it as steady as you can, in your typical voice, until I ask you to stop". The clinician may provide a model for this task. The speaker performs this task three times for 3-5 seconds each. "Next, you will say the sound of the vowel /i/. We will the do it as we have done for the vowel /a/. So you will hold it as steady as you can, in your typical voice, until I ask you to stop". The speaker performs this task three times for 3-5 seconds each.

2. Task 2: Sentences reading

The clinician should give to the speaker the six sentences printed on a paper.

The speaker should read progressively the sentences, one at a time. The clinician says, “Please read the following sentences one at the time, as if you were speaking to somebody in a real conversation”. The sentences are:

a. Num domingo esteve sol e fui com o avô António à explana “Évora” comer uma empada;
b. Segundo Simão, só Samuel sabe;
c. A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa;
d. É hora da Urraca ir à caça;
e. Onde eu brinco há um ninho de andorinhas encostado ao muro;
f. A Kika tapou a tua capa preta.

3. Task 3: Spontaneous speech

The clinician should ask the speaker to produce at least 20 seconds of natural conversational speech using a standard quote “Tell me about the place where you grew up”.

APPENDIX F: Application manual of II EP CAPE-V.

Escola Superior de Saúde do Instituo Politécnico de Setúbal

Instrumento de avaliação áudio-percetiva da voz:

"*Consensus Auditory-Perceptual Evaluation of Voice*"

2ª Versão Português-Europeu (II CAPE-V PE)

MANUAL DE APLICAÇÃO

Equipa de Investigação:

Sancha C. de Almeida

Ana Paula Mendes

2015

O "*Consensus Auditory-Perceptual Evaluation of Voice"* (CAPE-V) é um instrumento clínico de avaliação áudio-percetiva da voz. Este instrumento tem procedimentos específicos para recolha de amostras de voz e para avaliação das mesmas.

O CAPE-V utiliza as seguintes tarefas fonatórias: produção de vogais sustentadas, leitura de frases e produção de discurso espontâneo.

As **vogais sustentadas** selecionadas são [a, i] consideradas como vogais "relaxadas" e "tensas", respetivamente. Ambas as vogais são produzidas 3 vezes cada durante 3-5 segundos.

Para a **leitura de frases**, foram desenvolvidas seis frases com o objetivo de analisar diferentes comportamentos laríngeos e sinais clínicos:

a. Produção de todas as vogais orais e nasais do Português Europeu (PE) – "*Num domingo esteve sol e fui com o avô António à esplanada "Évora" comer uma empada*";
b. Ataques vocais suaves na transição de segmentos não vozeados para vozeados através de uma frase com palavras iniciadas /s/ - "*Segundo Simão, só Samuel sabe*";
c. Eventuais espasmos/bloqueios laríngeos através na produção todos os segmentos vozeados do PE– "*A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa*";
d. Ataque vocal forte nas palavras iniciadas por vogais– "*É hora da Urraca ir à caça*";
e. Hiponasalidade e possível estimulabilidade para a "*Resonant Voice Therapy*" através em todas as vogais e consoantes nasais do PE – "*Onde eu brinco há um ninho de andorinhas encostado ao muro*";
f. Hipernasalidade ou emissão de ar nasal através de frase composta por segmentos oclusivos não vozeados – "*A Kika tapou a tua capa preta*".

A produção de **discurso espontâneo** é elicitada pela questão "Fale-me do sítio onde cresceu". Esta tarefa tem a duração mínima de 20 segundos.

Os parâmetros da qualidade vocal analisados pelo CAPE-V são:

1. **Grau de severidade global:** Perceção global da alteração vocal;
2. **Rouquidão:** Irregularidade na fonte sonora percebida auditivamente;
3. **Soprosidade:** Escape de ar audível na voz;

4. **Tensão:** Perceção de esforço vocal excessivo (hiperfunção);
5. **Altura tonal:** Correlação percetiva com a frequência fundamental. Este parâmetro analisa se a altura tonal de um sujeito é muito desviante da altura tonal normal para um sujeito do mesmo sexo, idade e referencial cultural. A classificação (grave/agudo) deve ser indicada no espaço em branco por cima da escala, antes da marcação do desvio na linha;
6. **Intensidade:** Correlação percetiva com a intensidade sonora. Este parâmetro analisa se a intensidade vocal de um sujeito é muito desviante da intensidade vocal normal para um sujeito do mesmo sexo, idade e referencial cultural. A classificação (fraco/forte) deve ser indicada no espaço em branco por cima da escala, antes da marcação do desvio na linha.

Na folha de registo do CAPE-V, em frente a cada um dos seis parâmetros vocais encontra-se uma linha de 0-100 mm que forma uma escala visual análoga (EVA).

O juiz **deve indicar o grau de desvio da normalidade percebido auditivamente com um traço vertical sobre a escala** correspondente a cada um dos parâmetros. O juiz pode colocar um traço vertical em qualquer sítio ao longo da linha devendo o traço ser baseado nas observações diretas relativamente às características de cada voz.

Os extremos da linha da escala não são rotulados. Abaixo da linha da escala, encontram-se três categorias: desvio ligeiro (DL); desvio moderado (DM); e desvio severo (DS). Estas categorias indicam gradação da severidade do desvio e não a quantificação do desvio.

À direita de cada parâmetro vocal existem duas letras "C" e "I":

- "C" representa a consistência;
- "I" a inconsistência da presença de um parâmetro vocal particular.

O juiz **deve circular a letra que melhor descreve a consistência do parâmetro avaliado**. A avaliação de "consistente" indica que o parâmetro vocal esteve presente em todas as tarefas fonatórias. Contrariamente, a avaliação "inconsistente" indica que o parâmetro ocorreu de forma inconstante durante as diferentes tarefas fonatórias. Por exemplo, um sujeito pode exibir consistentemente uma qualidade vocal tensa ao longo de todas as tarefas fonatórias. Neste caso, o juiz deverá circular a letra "C". Contrariamente, outro sujeito pode exibir tensão constante durante a produção das vogais e inconsistente durante uma ou mais tarefas de fala encadeada. Neste caso, o juiz deverá circular a letra "I".

Na folha de registo do CAPE-V existem **duas escalas em branco** sem parâmetros atribuídos e características adicionais. O juiz deve utilizar estas duas **para avaliar parâmetros adicionais** que considere pertinentes para caracterização da voz em questão. No espaço "**factores adicionais"** o juiz **pode indicar a presença de outros atributos que não foram referidos anteriormente**. Por exemplo, se um sujeito estiver afónico, este facto deve ser registado no espaço "outros parâmetros" e não nas escalas sem parâmetros atribuídos. O juiz pode ainda indicar observações pertinentes acerca da ressonância na secção denominada "comentário sobre ressonância". Nesta secção podem ser incluídos comentários como por exemplo: "hipernasalidade", "hiponasalidade", "cul-de-sac", entre outros.

Antes de preencher a folha de registo do CAPE-V, o juiz deve observar o desempenho de cada sujeito ao longo todas as tarefas fonatórias e proceder a uma **análise global da qualidade vocal**. Deve ser preenchida uma folha de registo do CAPE-V por cada sujeito.

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

Juiz #_____

Amostra de voz #_____ **Data aplicação ___/___/___**

Os parâmetros da qualidade vocal devem ser medidos recorrendo às seguintes tarefas fonatórias:

1. **Vogais sustentadas /a/ e /i/** (três repetições de 3-5 segundos cada)
2. **Leitura de frases**:
   a. Num domingo esteve sol e fui com o avô António à esplanada "Évora" comer uma empada.
   b. Segundo Simão, só Samuel sabe.
   c. A Zé, mãe do Gabriel, deu-lhe um bolo de laranja e vinho velho de Runa.
   d. É hora da Urraca ir à caça.
   e. Onde eu brinco há um ninho de andorinhas encostado ao muro.
   f. A Kika tapou a tua capa preta.
3. **Discurso espontâneo** (mínimo 20 seg.) "Fale-me do sítio onde cresceu"

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

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

				Pontuação
Grau de severidade global	_____ C	I	____/100	
	DL DM DS			
Rouquidão	_____ C	I	____/100	
	DL DM DS			
Soprosidade	_____ C	I	____/100	
	DL DM DS			
Tensão	_____ C	I	____/100	
	DL DM DS			

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

_____ C	I	____/100	
DL DM DS			

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

_____ C	I	____/100	
DL DM DS			
_____ C	I	____/100	
DL DM DS			
_____ C	I	____/100	
DL DM DS			

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

_____

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

_____

APPENDIX G: Application manual of GRBAS.

Escola Superior de Saúde do Instituo Politécnico de Setúbal

Instrumento de avaliação áudio-percetiva da voz:

GRBAS

MANUAL DE APLICAÇÃO

Equipa de Investigação:

Sancha C. de Almeida

Ana Paula Mendes

2015

A escala GRBAS foi desenvolvida por Hirano (1981) para avaliação áudio-percetiva a qualidade vocal. Esta escala avalia os seguintes parâmetros vocais:

- "**G**" – Grau geral da alteração vocal;
- "**R**" – Rouquidão;
- "**B**" – Soprosidade;
- "**A**" – Astenia;
- "**S**" – Tensão.

A escala GRBAS não tem um protocolo de procedimentos de recolha de amostras de voz nem linhas orientadoras para a avaliação de cada um dos parâmetros vocais. A GRBAS usa uma escala de Likert de 4 pontos para avaliar a severidade de cada um dos parâmetros vocais sendo "0" normal, "1" alteração ligeira, "2" alteração moderada e "3" alteração severa.

Após ouvir todas as tarefas fonatórias (vogais sustentadas /a, i/, leitura de frases e discurso espontâneo) o juiz deve **avaliar os cinco parâmetros vocais da escala GRBAS.** Deve ser preenchida uma folha de registo da GRBAS por sujeito.

Juíz # _____ **Data de aplicação: ____ / _____/ _____**

Amostra de voz # _____

Classifique cada parâmetro vocal numa escala de "0" (normal), "1" (alteração ligeira), "2" (alteração moderada) e "3" (alteração severa).

Escala GRBAS[1]

G____ R____ B____A____S____

[1] Hirano (1981)

Legenda:

G = Grau

R = Rouquidão

B = Soprosidade

A = Astenia

S = Tensão

LIST OF REFERENCES

- Aaronson, N., Alonso, J., Burman, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. K (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research, 11*, 193 – 205.
- American Speech-Language-Hearing Association. (1997). *Guidelines for audiologic screening* [Guidelines]. Retrieved from http://www.asha.org/policy/GL1997-00199/.
- American Speech-Language-Hearing Association. (2006). *Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V).* Special Interest Division 3, Voice and Voice Disorders. Retrieved form http://www.asha.org/uploadedFiles/members/divs/D3CAPEVprocedures.pdf.
- Andy, F. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE.
- Awan, S. N., & Lawson, L. L. (2009). The effect of anchor modality on the reliability of vocal severity ratings. *Journal of Voice, 23*(3), 341 – 352.
- Barsties, B., & De Bodt, M. (2015). Assessment of voice quality: current state-of-the-art. *Auris, Nasus, Larynx, 42*(3), 183 – 188.
- Bassich, C. J., & Ludlow, C. L., (1986). The use of perceptual methods by new clinicians for assessing voice quality. *The Journal of Speech and Hearing Disorders, 51*(2), 123 – 133.
- Behlau M. (2004). Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), ASHA 2003 [Refletindo sobre o novo]. *Revista da Sociedade Brasileira de Fonoaudiologia, 9*, 187 – 189.
- Behlau, M., & Pontes, P. (1995). *A avaliação e tratamento das disfonias.* São Paulo: Lovise.
- Behrman, A. (2005). Common practices of voice therapists in the evaluation of patients. *Journal of Voice, 19*(9), 454 – 469.
- Bele, I. V. (2005). Reliability in perceptual analysis of voice quality. *Journal of Voice, 19*(4), 555 – 573.
- Bele, I. V. (2007). Dimensionality in voice quality. *Journal of Voice, 21*(3), 257 – 272.
- Bless, D. M., Baken, R. J., Hacki, T., Fritzell, B., Laver, J., Schutte, H., Hirano, M., Loebell, E., Titze, I., Faure, M. A., Muller, A. Wendler, J., Fex., S. Kotby, M.

N., Brewer, D., Sonninen, A., & Hurme, P. (1992). International association of Logopedics and Phoniatrics (IALP) voice committee discussion of assessment topics. *Journal of Voice*, 6(2), 194 – 210.

- Brinca, L., Batista, A. P., Tavares, A. I., Pinto, P. N., & Araújo, L. (2015). The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *Journal of Voice*, *29*(6), 776.e7 – 776.e14.
- Carding, P. N., & Mathieson, L. (2008). Voice and speech production. In M. Gleeson, G. G. Browning, M. J. Burton, R. Clarke, J. Hibbert, N. S. Jones, V. J. Lund, L. M., Luxon, J. C. Watkinson (Eds.), *Scott-Brown's Otorhinolaryngology, head and neck surgery* (pp. 2164-2169). London: Hodder Education.
- Carding, P. N., Carlson, E., Epstein, R., Mathieson, L., & Shewell, C. (2000). Formal perceptual evaluation of voice quality in the United Kingdom. *Logopedics Phoniatrics Vocology 25*(3), 133 – 138.
- Carding, P. N., Wilson, J.A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: state of the science review. *The Journal of Laryngology & Otology, 123*, 823 – 829.
- Chan, M. K., & Yiu, E. M-L. (2006). A comparison of two perceptual voice evaluation training programs for naïve listeners. *Journal of Voice, 20*(2), 229 – 241.
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine, 119*, 166.e7 – 116.e16.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Phychological Bulletin, 52*(4), 281 – 302.
- 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), 78 – 80.
- Dejonckere, P. H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., Van de Heyning, P., Remacle, M., & Woisard, V. (2001). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-Rhino-Laryngology, 258*, 77 – 82.

- Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de Laryngologie Otologie Rhinologie, 117*(3), 219-224.
- DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 155 – 164;
- Eadie, T. L, & Doyle, P. C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *Journal of Voice*, *19*(1), 1 – 14.
- Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgements of dysphonic voice. *Journal of Voice, 20*(4), 527 – 544.
- Eadie, T. L., Boven, L. V., Stubbs, K., & Giannini, E. (2010). The effect of musical background on judgements of dysphonia. *Journal of Voice, 24*(1), 93 – 101.
- Eadie, T., & Kapsner-Smith, M. (2011b). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research, 54*(2), 430 – 447.
- Eadie, T., Stroka, A., Wright, D. R., & Merati, A. (2011a). Does knowledge of medical diagnosis bias auditory-perceptual judgements of dysphonia? *Journal of Voice*, *25*(4), 420 – 429.
- Fex, S. (1992). Perceptual evaluation. *Journal of Voice, 6*(2), 155 – 158.
- Fortin, M. F. (1996). *O Processo de Investigação: Da concepção à realização*. Loures: Lusociência.
- Franic, D. M., Bramlett, R. E., & Bothe, A. C. (2005). Psychometric evaluation of disease specific quality of life instruments in voice disorders. *Journal of Voice, 19*(2), 300 – 315.
- Freitas, S. V., Pestana, P. M., Almeida, V., & Ferreira, A. (2014). Audio-perceptual evaluation of Portuguese voice disorders – an inter- and intrajudge reliability study. *Journal of Voice*, *28*(2), 210 – 215.

- Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards on voice quality judgments. *Journal of Speech and Hearing Research, 36*, 14 – 20.
- Ghio, A. Révis, J. Merienne, S., & Giovanni, A. (2013). Top-down mechanisms in dysphonia perception: the need for blind tests. *Journal of Voice, 27*(4), 481 – 485.
- Ghirardi, A. C., Ferreira, L. P., Giannini, S. P., & Latorre, M. R. (2013). Screening index for voice disorder (SIVD): development and validation. *Journal of Voice, 27*(2), 195 – 200.
- Gould, J., Waugh, J., Carding, P., & Drinnan, M. (2012). A new voice rating tool for clinical practice. *Journal of Voice, 26* (4), e163 – e170.
- Groove, F. L., & Shoyer, A. L. (2000). Clinical science research. *The Journal of Thoracic and Cardiovascular Surgery*, *119*, S11 – 21.
- Guimarães, I. (2007). *A Ciência e a Arte da Voz Humana*. Alcabideche: ESSA.
- Hammarberg, B. (2000). Voice Research and Clinical Needs. *Folia Phoniatrica et Logopaedica, 52*, 93 – 102.
- Harshbarger, T. R. (1977). *Introductory statistics: A decision map* (2nd ed.). New York: Macmillan.
- Helou, L. B., Solomon, N. P., Henry, L. R., Coppit, G. L., Howard, R. S., & Stojadinovic, A. (2010). The role of listener experience on consensus auditory-perceptual evaluation of voice (CAPE-V) ratings of postthyroidectomy voice. *American Journal of Speech-Language Pathology*, *19*, 248 – 258.
- Hirano, M. (1981). *Clinical examination of voice*. Vienna: Springer-Verlag.
- IBM SPSS, (2013). Statistical Package for the Social Sciences 22.0 for Windows [Computer software]. Armonk, NY: IBM Corp.
- Iwarsson, J., & Petersen, N. R. (2012). Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *Journal of Voice, 26* (3), 304 – 312.
- Jesus, L., Barney, A., Sá Couto, P., Vilarinho, H., & Correia, A. (2009b December). Voice Quality Evaluation Using CAPE-V and GRBAS in European Portuguese. In *poceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2009)*. Florence, Italy, pp. 61 – 64.

- Jesus, L., Barney, A., Santos, R., Caetano, J., Jorge J., & Sá Couto, P. (2009a, September). Universidade de Aveiro's voice evaluation protocol. *In Proceedings of InterSpeech.* Brighton, UK, pp. 971 – 974.
- Jöreskog, K.G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.
- 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.
- Kelchner, L. N., Brehm, S. B., Weinrich, B., Middendorf, J., deAlarcon, A., Levin, L., & Elluru, R. (2010). Perceptual evaluation of severe pediatric voice disorders: rater reliability using consensus auditory perceptual evaluation of voice. *Journal of Voice, 24* (4), 441 – 449.
- Kelly, P. A., O'Malley, K. J., Kallen, M. A., & Ford, M. E. (2005). Integrating validity theory with use of measurement instruments in clinical settings. *Health Services Research, 40*(5, part II), 1605-1619;
- Kempster, G. B., Guerratt, 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,* 124 – 132.
- Kimberlin, C. L., & Winterstein, Al. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65 (1), 2276 – 2284.
- Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measurements of voice quality. *The Journal of the Acoustical Society of America, 104*(3), 1598 – 1608.
- Kreiman, J., & Gerratt, B. R. (2011). Comparing two methods for reducing variability on voice quality measurements. *Journal of Speech-Language and Hearing Research, 54*(3), 803 – 812.
- Kreiman, J., Gerratt, B. R., & Berke, G. S. (1994). The multidimensional nature of pathological vocal quality. *The Journal of the Acoustical Society of America, 96*(3), 1291 – 1302.

- Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33*, 103 – 115.
- Kreiman, J., Gerratt, B. R., Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122*(4), 2354 – 2356.
- Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36* (1), 21 – 40.
- Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*, 512 – 520.
- Kreiman, J., Vanlancker-Sidtis, D., Gerratt, B. (2004, June). Defining and measuring voice quality. In: *Proceedings from Sound to Sense* (MIT). Cambridge, Massachusetts, USA, pp. C-163 – C-168.
- Laver J., Wirz S., MacKenzie J., Hiller S. (1981). *A perceptual protocol for the analysis of vocal profiles*. Edinburgh: University of Edinburgh (Department of Linguistics).
- Law, T., Kim, J. H., Lee, K. Y., Tang, E.C., Lam, J. H., Van Hasselt, A. C., & Tong, M. C. (2012). Comparison of rater's reliability on perceptual evaluation of different types of voice sample. *Journal of Voice, 26*(5), 666.e13 – 666.e21.
- Leong, K., Hawkshaw, M. J., Dentchev, D., Gupta, R., Lurie, D., & Sataloff, R. T. (2013). Reliability of objective voice measures of normal speaking voices. *Journal of Voice, 27* (2), 170 – 176.
- Lohr, K. N., Aaronson, N. K., Alonso, J., Burman, M. A., Patrick, D. L., Perrin, E. B., & Roberts, J. S. (1996). Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clinical Therapeutics, 18*(5), 979 – 992.
- Maryn, Y., & Roy, N. (2012). Sustained vowels and continuous speech in the audioty-perceptual evaluation of dysphonia severity. *Jornal da Sociedade Brasileira de Fonoaudiologia, 24*(2), 107 – 112.

- McAlliser, A., Sundberg, J., & Hibi, S. R. (1996). Acoustic measurements and perceptual evaluation of hoarseness in children's voices. *Speech, Music, and Hearing – Quarterly Progress and Status Repoort, 37*(4), 15 – 26.
- McBurney, D. H., & White, T. L. (2007). *Research Methods, Seventh Edition*. Belmont: Thomson Wadsworth.
- McGlashan, J., & Fourcin, A. (2008). Objective evaluation of the voice. In M. Gleeson (Ed.), *Scott-browns otorhynolaryngology, head and neck surgery* (pp. 2170-2191). London: Hodder Arnold Publishers.
- Mehta, D. D., & Hillman, R. E. (2008). Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Current Opinion in Otolaryngology & Head and Neck Surgery, 16*(3), 211 – 215.
- Moerman, M. B. J., Martens, J. P., Crevier-Buchman, L., de Haan, E., Grand, S., Tessier, C., Woisard, V., & Dejonckere, P. H. (2006a). Perceptual evaluation of substitution voices: development and evaluation of the (I)INFVo rating scale. *European Archives of Oto-Rhino-Laryngology and Head & Neck, 263*, 435 – 439;
- Moerman, M. B. J., Martens, J. P., Van der Borgt, M. J., Peleman, M., Gillis, M., & Dejonckere, P. H. (2006b). Perceptual evaluation of substitution voices: development and evaluation of the (I)INFVo rating scale. *European Archives of Oto-Rhino-Laryngology and Head & Neck, 263*, 183 – 187;
- 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 Logopeadica*, *65*(5), 257 – 65.
- Nemr, K., & Lehn, C. (2010). Voz em Câncer de Cabeça e Pescoço. In Fernandes, F. et al. (2nd ed.) "*Tratado de Fonoaudiologia*" (pp. 798). São Paulo: Roca.
- Nemr, K., Simões-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *Journal of Voice, 26*(6), 812.e17 – 812.e22.
- Nemr, K., Simões-Zenari, M., Souza, G. G., Hachiya, A., & Tsuji, D. H. (2015). Correlation of the dysphonia severity index (DSI), consensus auditory-perceptual evaluation of voice (CAPE-V), and gender in Brazianls with and without voice disorders. *Journal of Voice*, *25*.

- Núñez-Batalla, F. Morato-Galán, M., García-López, I., & Ávila-Menéndez, A. (2015). Validation of the Spanish adaptation of the consensos auditoy-perceptual evaluation of voice. *Acta otorrinolaringológica Española, 66*(5), 249 – 257.
- Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica, 61*, 49 – 56.
- Orlikoff, R. F. (1999). The perceived role of voice perception in clinical practice. *Phonoscope, 2*(2), 87 – 106.
- Patel, S., & Shrivastav, R. (2007). Perception of dysphonic vocal quality: some thoughts and research update. *ASHA SIG 3 Perspectives on Voice and Voice Disorders, 17*, 3 – 7.
- Patel, S., Shrivastav, R., & Eddins, D. A. (2010). Perceptual distances of breathy voice quality: a comparison of psychophysical methods. . *Journal of Voice*, 24(2), 168 – 177.
- Pinho, S.M.R. & Pontes, P. (2008). *Músculos intrínsecos da Laringe e Dinâmica Vocal*. (Série Desvendando os Segredos da Voz). (Vol. 1). Rio de Janeiro: Revinter.
- Pinho, S.M.R., & Pontes, P. (2002). Escala de Avaliação Perceptiva da Fonte Glótica: RASAT. *Vox Brasilis, 3*, 11-13.
- Rabinov, C. R., Kreiman, J., Gerratt, B. R., & Bielamowicz, S. (1995). Comparing reliablity of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech and Hearing Research, 38*, 26 – 32.
- Royal College of Speech & Language Therapists (2009). RCSLT resource manual for commissioning and planning services for SLCN, voice. Oxon: British Library Cataloguing. Retrieved from http://www.rcslt.org/speech_and_language_therapy/commissioning/voice_plus_intro
- Sáenz-Lechón, N., Godino-Llorente, J. I., Osma-Ruiz, V., Blanco-Velasco, & M., Cruz-Roldán, F. (2006, September). Automatic assessment of voice quality according to the GRBAS scale. In *Conference Engineering in Medicine and Biology Society*, 2006. *(EMBS 2006)*. New York, US, pp. 2478 – 2481.
- Schwartz, S. R., Cohen, S. M., Dailey, S. H., Rosenfeld, R. M., Deutsch, E. S., Gillespie, M. B., Granieri, E., Hapnet, E. R., Kimball, C. E., Krouse, H. J., McMurray, J. S., Medina, S., O'Brien, K., Ouellette, D. R., Messinger-Rapport,

B. J., Stachler, R. J., Strode, S., Thompson, D. M., Stemple, J. C., Willging, J. P., Cowley, T., McCoy, S., Bernad, P. G., & Patel, M. M. (2009). Clinical practice guideline: Hoarseness (Dysphonia). *Otolaryngology Head and Neck Surgery*, 141, S1-S31;

- Shewell, C. (1998). The effect of perceptual training on ability to use the vocal profile analysis scheme. *International Journal of Language & Communication Disrorder, 33(S1)*, 222-226.
- Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory of measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research, 48*, 323-335.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420 – 428.
- Sofranko, J. L., & Prosek, R. A. (2012). The effect of experience on classification of voice quality. *Journal of Voice*, *26*(3), 299 – 303.
- Solomon, N. P., Helou, L. B., & Stojadinovic, A. (2011). Clinical versus laboratory ratings of voice using the CAPE-V. *Journal of Voice*, *25*(1), e7 – e14.
- Speyer, R. (2008). Effects of voice therapy: a systematic review. *Journal of Voice*, *22*(5), 565 – 580.
- Verdolini, K., Rosen, C. A., & Branski, R. C. (2006). *Classification Manual for Voice Disorders – I*. New Jersey: Lawrence Erlbaum Associates Publishers.
- Vilelas, J. (2009). *Investigação – O Processo de Construção do Conhecimento*. Lisbon: Edições Sílabo.
- 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 Otorhinolarygology*, *261*(8), 429 – 434.
- Wilson, D. (1987). *Voice problems of children*. Baltimore: Williams and Wilkins.
- 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. *Journal of Voice*, 13(4), 508 – 517.
- Zraick, R. I., Kempster, G. B., Connor, N. P., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E. (2011). Establishing validity of the consensus auditory-

perceptual evaluation (CAPE-V). *American Journal of Speech-Language Pathology*, *20*(1), 14 – 22.

- Zraick, R. I., Wendel, K., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice*, 19(4), 574 – 581.