THE EFFECTS OF CLINICAL EXPERIENCE ON THE VARIABILITY AND RELIABILITY OF CAPE-V RATINGS IN NON-PATHOLOGICAL VOICES

By

Anthony Joseph Strevett

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Communicative Sciences and Disorders – Master of Arts

2021

ABSTRACT

THE EFFECTS OF CLINICAL EXPERIENCE ON THE VARIABILITY AND RELIABILITY OF CAPE-V RATINGS IN NON-PATHOLOGICAL VOICES

By

Anthony Joseph Strevett

This study sought to address limitations in the current literature by studying the effects of years of clinical experience on the variability and reliability of ratings using the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) (Kempster et al., 2009). This study compared the ratings of a group of inexperienced speech-language pathologists to those of a group of experienced speech-language pathologists with extensive background in voice disorders. This study used voice recordings of non-therapy seeking individuals to provide non-pathological voice samples for the two groups of clinicians to rate. This was done to address the paucity of research focusing on the reliability of a subset range of the CAPE-V. It was hypothesized that the inexperienced clinician will demonstrate greater variability and less reliability in using the CAPE-V to rate non-pathological voices. The resulting data supports these hypotheses generally, though inferential statistics were ineffective methods of analysis due to limitations of study.

Copyright by
ANTHONY JOSEPH STREVETT
2021

To Mom.
Hang tight, Papa.

ACKNOWLEDGEMENTS

I would like to acknowledge several individuals for their dedication to my success.

First, the members of my thesis committee, Dr. Dimitar Deliyski, Dr. Maryam Naghilbolhosseini, Dr. Jeff Searl, and Dr. Bridget Walsh, all of whom provided amazing feedback, guidance, and encouragement along the way. Thank you for contributing to my growth as a researcher and clinician.

Second, Dr. David Ford, who played an essential advisory role in this project.

Third, Kristin Hicks, Leslie Fernandez-Lopez, and Kim Winkle, who always have a smile on their face and a little piece of happiness to share.

And lastly, my parents, Keith and Stacy Strevett, who are so willing to help and encourage, even when they don't know what I'm talking about.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

LITERATURE REVIEW

One aspect of the speech-language pathologist's job is caring for the voices of his or her patients. A speech-language pathologist (SLP) has many diagnostic tools available to them including instrumental evaluation of the larynx, acoustic and aerodynamic analysis, and patient assessment of vocal handicap. Another tool available for use is the auditory-perceptual assessment of voice.

Auditory-perceptual judgements are the most used diagnostic element by speech-language pathologists (Behrman, 2005; Eadie & Doyle, 2005). To use this method, clinicians listen to a client's voice and rate different vocal characteristics on a scale (Awan & Lawson, 2009). This is a subjective measurement and relies on the clinician's ability to accurately perceive what is being produced by the client's vocal tract (Bele, 2005). Auditory-perceptual judgement allows the clinician to distinguish between normal and pathological voices, to determine the severity of the pathology if one exists, and to plan and adjust treatment goals (Bassich & Ludlow, 1986; Yamasaki et al., 2017). It is cited as the gold standard for voice evaluations (Chan & Yiu, 2002; Helou et al., 2010). This current study focuses on the variability and reliability of these ratings as a function of the level of experience of the clinician.

Types of Auditory-Perceptual Assessment Scales

The auditory-perceptual assessment scales allow the clinician to make observations on various vocal characteristics. These scales are typically ordinal or visual analog. Ordinal scales provide clinicians with categorical options with a natural order (e.g., none, mild, moderate, and high) and the clinician selects one of these options to describe the vocal characteristic being

judged. These limited options increase the intra-rater and inter-rater reliability but reduce the ability to specifically represent vocal characteristics (Yu et al., 2002).

Visual analog (VA) scales are typically 100-millimeter lines on which clinicians place marks based on the perceived severity of a vocal characteristic (Awan, 2009). Marks closer to one end of the line represent less severe deviancy for a given voice characteristic; marks closer to the other edge represent greater severity (Awan, 2009). Contrary to ordinal scales, VA scales increase the ability to specifically represent characteristics of the voice but reduce the intra-rater and inter-rater reliability (Kreiman et al., 1993). The reasons for this are discussed later in this literature review.

Examples of Auditory-Perceptual Assessments

Many different auditory-perceptual assessments are available, including the GRBAS scale (Hirano, 1981), the Buffalo III Voice Profile (Wilson, 1987), and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) (Kempster et al, 2009). The GRBAS scale (Hirano, 1981) includes ratings of the following vocal characteristics: grade (overall severity of voice deviance), roughness, breathiness, asthenia (vocal weakness and propensity to fatigue), and strain. For further definition of these characteristics, please see Appendix A. The GRBAS scale requires the clinician to rate each of these five vocal characteristics on a four-point scale: “Normal”, “Slight”, “Moderate”, or “Extreme” (Hirano, 1981). In that sense, then, the GRBAS is an ordinal scale.

The Buffalo III Voice Profile (Wilson, 1987) analyzes the following vocal characteristics: pitch, intensity (loudness), quality, resonance, and adverse social/emotional/academic/vocational effects. Clinicians rate each of these vocal characteristics on a five-point scale: “Normal”,

"Mild", "Moderate", "Severe", and "Very Severe" (Wilson, 1987). Like the GRBAS, the Buffalo III Voice Profile is an ordinal scale. A copy of this assessment has been attached as Appendix B.

The CAPE-V (Kempster et al., 2009) is similar to the two previously mentioned assessments in that it is an auditory-perceptual assessment. However, it employs a VA scale. This distinction is important because it allows a clinician to represent with greater detail the perceived deviance in a vocal characteristic. The CAPE-V measures six characteristics of the voice: overall severity, roughness, breathiness, strain, pitch, and loudness (Kempster et al., 2009). According to the authors of the CAPE-V, these six features are the most common to appear in descriptions of voice across scholarly literature. As noted above, the clinician places a mark along the VA scale line to represent their perception of a vocal parameter. The distance from the left edge of the line to the mark is measured and reported as the rating for that vocal parameter. For example, if the mark is 73 mm from the left edge of the line, the rating for that particular vocal characteristic is 73. To supplement this numerical rating, clinical descriptors (e.g., mild, moderate-severe, etc.) are provided with the VA scale of the CAPE-V to augment the ratings. The clinician may choose to report these clinical descriptors as well, at their discretion.

To administer the CAPE-V, clinicians follow a protocol, eliciting specified voice samples from the client (Kempster et al., 2009). These voice samples include vowel phonations, reading of specific sentences, and a spontaneous response to a specified prompt (Kempster et al., 2009). A copy of the CAPE-V assessment has been attached as Appendix C.

The CAPE-V resulted from a need for a standardized assessment that incorporates the ability to specifically represent vocal characteristics that a VA scale provides. A standardized assessment should be administered and scored the same way regardless of patient, clinician, and setting (Fex, 1992). The CAPE-V was developed with this in mind in an attempt to improve the

evaluation and documentation of voice quality (Kempster et al., 2009). The CAPE-V allows for data which is "sound theoretically, is clinically meaningful, and can be consistently measured," (Kempster et al., 2009, p. 125). It has since become a widely used auditory-perceptual assessment for speech-language pathologists (Helou et al., 2010). Because the CAPE-V is widely used in clinical and research activities, it is the specific focus of this study.

Reliability of the CAPE-V

The relative merit of an assessment – which has implications on its usefulness to the profession – depends, in part, upon the reliability of the tool. Reliability refers to the level at which scores of a scale are the same across multiple administrations of that assessment (Zraick et al., 2011). There are two distinct types of reliability: intra-rater and inter-rater. Both have important implications on the certainty with which clinicians use assessments, and several studies have been conducted to research these for the CAPE-V.

Intra-rater reliability is the degree to which the same clinician rates the same vocal characteristic similarly, across multiple administrations of the assessment. High intra-rater reliability is important for assessments because it demonstrates that its use is not dependent on situational variables. Inter-rater reliability is the degree to which different clinicians (i.e., "raters") agree on a score and rate a vocal characteristic similarly. High inter-rater reliability is important for assessments because it demonstrates that its use is not dependent (at least not to a great extent) on the idiosyncrasies of the clinicians.

Karnell et al. (2007) studied intra- and inter-rater reliability of four experienced speech-language pathologists using the CAPE-V. These individuals were specialists in the field of voice, and one was the clinician who originally rated the voice samples used in the study (Karnell et al., 2007). Prior to listening to the pathological voice samples, the clinicians listened to four voice

samples: a female and a male without voice pathology, and a female and a male with voice pathology rated as highly severe on the overall severity dimension. This was done to establish external anchors with agreed-upon ratings to which clinicians can compare all subsequent voice samples in the experiment. 34 pathological voice samples were rated for each CAPE-V dimension by the four clinicians. This occurred twice, with the two rating sessions separated by seven days. This repeated listening of all 34 voice samples was used to analyze intra-rater reliability. The researchers used Spearman's correlation coefficient to determine intra-rater reliability and reported a range of 0.88 to 0.97, which allowed them to conclude that the intra-rater agreement was acceptable. The inter-rater reliability was determined with Spearman's correlation coefficient as well, and the researchers reported a range of 0.86 to 0.93 and the authors concluded that the CAPE-V had acceptable inter-rater agreement.

Kelchner et al. (2010) performed a similar study on pathological pediatric voices and found strong intra- and inter-rater reliability as well. Three speech-language pathologists, who specialized in the field of voice and had at least seven years of experience, served as raters (Kelchner et al., 2010). Each clinician rated 50 voice samples using the CAPE-V initially and, after seven days, listened to and rated 17 of the original voice recordings to allow for the analysis of intra-rater reliability. The researchers used intraclass correlation coefficients (ICC) to determine the degree of inter-rater reliability for each of the six vocal characteristics measured by the CAPE-V. The researchers reported moderately strong agreement in the characteristics of overall severity (0.67), roughness (0.68), breathiness (0.71), and pitch (0.68). Fair agreement was found for strain (0.35) and loudness (0.57). ICC was also used to analyze intra-rater reliability, and the researchers reported an average of 0.87 for overall severity, 0.82 for roughness, 0.82 for

breathiness, 0.63 for strain, 0.78 for pitch, and 0.79 for loudness. The authors concluded that the ICC results indicate moderately strong intra-rater agreement.

Nemr et al. (2012) investigated intra- and inter-rater reliability and found results similar to the previous two studies. The authors recruited three speech-language pathologists, each with more than five years of experience, to listen to and rate voice samples using the CAPE-V (Nemr et al., 2012). 60 voice samples were used, 50 of which were of pathological voices, as determined by clinicians at a local medical center. The remaining 10 were of non-pathological voices and served as a control samples. Each clinician completed the CAPE-V for all 60 voice samples, with an additional six randomly repeated voice samples to allow for analysis of intra-rater reliability. The researchers used ICC to calculate intra- and inter-rater reliability and found strong agreement in both. Intra-rater ICC ranged between 0.927 and 0.985. Inter-rater ICC averages were reported for each vocal characteristic as follows: overall severity (0.911), roughness (0.870), breathiness (0.897), and strain (0.828). The researchers did not discuss separately the ratings given to the non-pathological voices.

Factors Affecting Reliability

Several articles have identified factors that influence CAPE-V ratings, thereby affecting intra- and inter-rater reliability. Some of the identified factors include setting (e.g., public schools versus outpatient clinics) (Solomon et al., 2011), presence of external voice anchors (provided examples of pathological and non-pathological voice samples) (Awan & Lawson, 2009; Eadie & Kaspner-Smith, 2011; Kelchner et al., 2010; Yiu et al., 2007), psychological predispositions (Kent, 1996; Warren, 1976), and rater experience (Bassich & Ludlow, 1986; Eadie & Baylor, 2006; Helou et al., 2010). This present study will focus on the last factor: experience.

Effects of Psychology on Reliability

One influence on reliability that is closely related to experience is that of the psychological makeup of the clinician. This includes the biases that the clinician may hold during the assessment, based on what they expect to hear. Warren (1967) stated that the hypotheses developed by the listener to interpret nascent auditory and phonemic signals play a significant role in the interpretation of these signals (Warren, 1967). Sometimes, the interpretation is not the intended message. For example, Warren reports that listeners, when exposed to a recording of a single word repeated many times consecutively, reported having heard several different words (Warren, 1967). This is because the listener's perception of the incoming speech signal changed, though the stimulus did not.

Kent (1996) notes that a similar phenomenon could occur while a clinician assesses a potentially pathological voice. He writes that "knowledge-based hypotheses are pervasive influences" (Kent, 1996, p. 8). Clinicians may misinterpret the auditory signals of the voice, given the perceptual nature of the assessment. This is an interesting influencing factor to note, as this present study will be using voice samples of clients who are non-therapy seeking.

Effects of Experience on Reliability

Experience has been shown to improve reliability in auditory-perceptual voice assessments (Bassich & Ludlow, 1986; Eadie & Baylor, 2006; Helou et al., 2010). One of the mechanisms behind this is the development of internal standards. These are a perceptual reference that the clinician draws upon when assessing the voice (Kreiman et al., 1993). Clinicians develop internal standards over years of clinical work, having listened to many voices – both pathological and non-pathological. In that sense, then, experienced clinicians have a broader base of comparison when judging severity of different vocal characteristics compared to

less experienced clinicians. The studies described below have all investigated the extent to which experience and internal standards can influence the reliability of auditory-perceptual voice assessment ratings.

Bassich and Ludlow (1986) conducted a multi-step experiment on how experience (specifically in the form of training) influences reliability in voice assessment. The researchers recruited four individuals with limited clinical experience, three of whom were speech-language pathology graduate students, and one was a vocal performance student. The researchers created a 4-point scale for pitch stability and loudness stability, a 5-point scale for pitch breaks, tremor, waver, rough-fry, wet-hoarse, harsh-shrill, breathy, strain-strangle, and nasality, and a 7-point scale for pitch level and loudness level. The four raters were then given descriptions of each of these dimensions and asked to rate one non-pathological and four pathological voice samples. Percent agreement was calculated for inter-rater reliability for each of the 13, and all but two dimensions were found to be below chance agreement. The four raters then completed 16 half-hour training sessions to increase inter-rater reliability. Then the four raters were asked to rate 10 non-pathological voices and 10 pathological voices. An ICC was calculated for each of the 13 dimensions, for both the non-pathological control samples and the pathological samples. In the control samples, six dimensions had >0.80 interrater reliability, and in the experimental samples, five dimensions had >0.80 interrater reliability. As a result, Bassich and Ludlow suggested that using an auditory-perceptual voice assessment requires "professional experience and sophistication" (Bassich & Ludlow, 1986, p. 131). They did not compare a group of experienced clinicians to a group of inexperienced clinicians. They also did not use the CAPE-V, which will be employed in this present study.

Eadie and Baylor (2006) conducted a similar study, looking at the effects of explicit training on the reliability of voice ratings. The researchers recruited 16 inexperienced raters who were graduate students in speech-language pathology. Prior to training, the students rated the overall severity, breathiness, and roughness of 30 adult pathological and 6 adult non-pathological voice samples using a 100-mm VA scale. After this baseline was obtained, each member of the group received two hour-long training sessions provided by the researchers. During this training, descriptions of vocal characteristics being rated were given. External anchors were also provided via recordings. The listeners then participated in a post-training session to rate the voice samples presented in a different order from the original. Mean Pearson coefficient correlations were calculated for both intra- and inter-rater reliability. Intra-rater reliability was reported for each vocal characteristic: overall severity, mean $r = 0.922$ (pre) and 0.961 (post); breathiness, mean $r = 0.698$ (pre) and 0.807 (post); and roughness, mean $r = 0.794$ (pre) and 0.830 (post). Inter-rater reliability was reported similarly: overall severity, mean $r = 0.862$ (pre-training) and 0.871 (post-training).; breathiness, mean $r = 0.616$ (pre) and 0.796 (post); and roughness, mean $r = 0.700$ (pre) and 0.691 (post) (Eadie & Baylor, 2006). The researchers concluded that overall severity was the vocal characteristic most reliably rated, and that training does improve intra- and inter-rater agreement. This study did not include a control group of raters who did not complete the two hours of training. Also, this study did not directly compare the ratings of inexperienced clinicians to those of experienced clinicians; nor did the study utilize a standardized assessment protocol such as that of the CAPE-V. The current study addressed both of these issues.

Helou et al. (2010) performed a study comparing CAPE-V ratings of inexperienced clinicians to those of experienced clinicians. Two groups of 10 individuals participated: an experienced clinician group included five speech-language pathologists and five

otolaryngologists; and an inexperienced clinician group consisted of ten medical professionals with no background or experience in the field of voice disorders. Both groups were provided a brief overview on using the CAPE-V and were given six voice samples to serve as external anchors. The two groups rated voice samples from patients who had recently undergone thyroidectomies. The voice samples had vocal characteristics ranging from normal to severely dysphonic, as determined by consensus rating of the authors. The two groups of clinicians were instructed to listen to each voice sample twice, and to rate each voice sample using the CAPE-V. Variability was analyzed by subtracting each listener's rating from the average rating and squaring the difference. The scores were reported as follows for the experienced and inexperienced raters, respectively: overall severity, 74.7 and 221.34; roughness, 77.1 and 230.7; breathiness, 97.1 and 329.4; and strain, 105.9 and 297.9 (Helou et al., 2010). The researchers determined that there was a significantly less variability for the experienced compared to the inexperienced group. ICC was used to calculate intra- and inter-rater reliability for each vocal characteristic (Helou et al., 2010). Intra-rater ICC were reported as follows for the experienced and inexperienced raters, respectively: overall severity, 0.911 and 0.838; roughness, 0.866 and 0.799; breathiness, 0.873 and 0.794; and strain, 0.793 and 0.582. Using the Mann-Whitney *U* test, the researchers found that the intra-rater reliability did not differ significantly between the two groups. Inter-rater ICC were reported as follows for the experienced and inexperienced raters, respectively: overall severity, 0.722 and 0.528 roughness, 0.636 and 0.566; breathiness, 0.625 and 0.415; and strain, 0.309 and 0.271 (Helou et al., 2010). Using methodology described by Chan and Yiu (2002), the authors found that the inter-rater reliability differed significantly between the two groups with greater reliability evidenced by the experienced group. The study

was limited in that it did not include non-pathological voices in the samples. This present study will seek to address this limitation.

Purpose of Present Study

As can be seen from this review of current literature, no studies have investigated the effects of experience on the reliability of CAPE-V ratings by comparing inexperienced clinicians to experienced clinicians with the sole focus on non-pathological voice samples. Various studies have investigated these factors peripherally (Nemr et al, 2012; Bassich & Ludlow, 1986; Eadie & Baylor, 2006; Helou et al., 2010). Nemr et al. (2012) did not separately analyze the ratings of the non-pathological voice samples. Bassich and Ludlow (1986) did not compare inexperienced clinicians' ratings to experienced clinicians' ratings. The study by Eadie and Baylor (2006) was limited in the same way. Helou et al. (2010) did not include non-pathological voices in the pool of voice samples. This present study seeks to address all these limitations.

This study is of clinical and theoretical importance because the CAPE-V allows for the scoring of normal (non-pathological) to mildly pathological voices. As such, research must be conducted to test the reliability of this portion of the assessment. Many studies have been completed on the reliability and variability of ratings using voice samples with moderate to severe dysphonia. This present study will add to the literature through the investigation of the reliability of the CAPE-V in non-pathological voices.

The purpose of this study is to learn about the impact of clinician experience on the variability and reliability of perceptual ratings of overall severity, roughness, breathiness, and strain. The expectation is that inexperienced clinicians will display greater variability and demonstrate less reliability in their ratings than the experienced clinicians. The detailed research questions and hypotheses are listed below:

RQ1: Does the experience of speech-language pathologists affect the variability of perceptual ratings?

H1: The inexperienced clinicians will demonstrate greater variability in rating overall severity than experienced clinicians.

H2: The inexperienced clinicians will demonstrate greater variability in rating roughness than experienced clinicians.

H3: The inexperienced clinicians will demonstrate greater variability in rating breathiness than experienced clinicians.

H4: The inexperienced clinicians will demonstrate greater variability in rating strain than experienced clinicians.

RQ2: Does the experience of speech-language pathologists affect the reliability of perceptual ratings?

H5: The inexperienced clinicians will demonstrate less intra- and inter-rater reliability in rating overall severity than experienced clinicians.

H6: The inexperienced clinicians will demonstrate less intra- and inter-rater reliability in rating roughness than experienced clinicians.

H7: The inexperienced clinicians will demonstrate less intra- and inter-rater reliability in rating breathiness than experienced clinicians.

H8: The inexperienced clinicians will demonstrate less intra- and inter-rater reliability in rating strain than experienced clinicians.

CHAPTER 2:

METHODS

Voice Samples

The voice samples used in this study were part of a pre-existing de-identified dataset resulting from a previous project. 36 voice samples were used in this study, from a corpus of 46 voice samples compiled for a doctoral dissertation project that was approved by the Institutional Review Board (IRB) at Michigan State University, (Study#: 00004971). 10 samples were excluded from the original dissertation project due to deviations from the protocol during collection. The voice samples were collected remotely via participant-operated instruments with the assistance of study team members via teleconference. Participants followed written and verbal instructions on a secured Zoom meeting to setup and record voice samples.

Voice samples were collected with an AKG P170 cardioid condenser microphone which was connected to an audio recording interface (Focusrite Scarlett Solo, 3rd generation). The microphone was placed on a Gator Frameworks GFW-MIC-0250 microphone stand and positioned 20 cm from the participant's mouth. The recordings were made with the participants' computer using Audacity software (version 2.4.1). The recordings were saved as .WAV files and submitted to the researcher for analysis via the postal service. All participants providing voice samples signed a written consent form allowing research on human subjects in accordance with Michigan State University's IRB requirements.

Participants recorded themselves producing sustained phonation of /a/ and /i/ vowels, reading of six sentences modeled after standard CAPE-V protocol detailed by Kempster et al. (2009) (see CAPE-V form in Appendix C), and reading the Rainbow Passage (Fairbanks, 1960) (see Appendix D). Participants were instructed to vocalize and read at habitual pitch and

loudness in a quiet environment. For this study, only the sustained vowel phonation and the Rainbow Passage were used. Average length of sustained phonation recordings was five seconds and average length of Rainbow Passage recordings was 38 seconds. These recordings were clipped to isolate portions during which the participants were phonating or reading.

Inclusion criteria for individuals providing voice samples required them to be in generally good medical condition and not be actively receiving voice or speech therapy. Participants were also screened for other conditions that could adversely affect speech production, such as breathing difficulties and hearing impairments. This was to ensure that all individuals were non-therapy seeking (NTS) and to exclude participants with overtly rough, breathy, or strained voices. Other inclusion criteria required participants to be between the ages of 18 and 70 years. The average age of the participants was 39.9 years with a standard deviation of 14.6 years. Both male and female voice samples were collected, with 47.2% of the participants being male and 52.8% being female. Table 1 displays the distribution of age and gender. All participants spoke fluent English.

Table 1:

Demographics of individuals providing voice samples

Age (in years)	18-31	32-44	45-57	58-70
Males	6	5	3	3
Females	6	6	4	3
Total	12	11	7	6

Listeners

The ratings of the voice samples were done by two groups of listeners that differed in their experience level as SLPs. The inexperienced group was comprised of three recently graduated SLP recruited via convenience sampling. This included one male and two females, between the age of 25 and 30 years, who had graduated approximately three months prior with an M.A. in Communicative Sciences and Disorders from Michigan State University. They had completed a 3-credit hour graduate course on Voice Disorders as a required component of their graduate program. During the course, they were asked to complete a limited number of rating assignments using the CAPE-V. They had less than 18 months clinical experience as a student and none had a sub-specialization in voice disorders. They completed their graduate program with above passing grades and their hearing was within normal limits.

Three listeners comprised the experienced clinician group. These individuals were three males who had 10 to 30 years of clinical experience working with people who have a voice disorder. Their ages were between 35 and 65 years. These individuals were recruited via convenience sampling. The ratings generated by this group were part of previously collected dataset.

All listeners completed an orientation via a study-specific training video to calibrate their use of the CAPE-V by establishing internal anchors. This has been shown to increase reliability of listeners (Eadie & Kaspner-Smith, 2011; Kelchner et al., 2010). All participants providing ratings signed a written consent form allowing research on human subjects in accordance with Michigan State University's IRB requirements.

Auditory-Perceptual Assessment

The CAPE-V was used in this study by the listeners to rate the voice samples. The standard CAPE-V protocol solicits ratings for overall severity, roughness, breathiness, strain, pitch, and loudness (Kempster et al., 2009). However, this study was only concerned with the deviation of overall severity, roughness, breathiness, and strain of the voice. As such, ratings for the other two dimensions were not obtained. This decision was made based on two factors. First, most studies have found that reliability is greatest for these four vocal characteristics (De Bodt et al., 1997; Webb et al., 2004; Kelchner et al., 2010). Second, perceived deviance from the norm in loudness and pitch are more attributable to age, gender, and upbringing than are the other four vocal characteristics assessed by the CAPE-V (Kelchner et al., 2010). Any references hereafter to the CAPE-V will refer to this modified version used in this study.

Procedure

Both groups of listeners were first orientated to the task by watching a training video. The video instructed the listeners in using the CAPE-V through a brief overview of using a VA scale and defining the four vocal characteristics being rated. The video also provided internal anchors for what constitutes a pathological voice by providing examples of voices that clinicians have deemed pathological. Establishing these internal anchors decreases the number of biasing factors through standardization of usage and clarification of the expectations of listeners. The listeners were then given access to the voice samples and asked to rate the voice samples using the CAPE-V along the four vocal characteristics being studied. Listeners completed their ratings at home, in a room free from distractions. They listened to the voice samples through QuickTime Player as .WAV files, via headphones or earphones. Listeners were instructed to play voice samples at a loudness that roughly mimicked the decibels of a normal conversation (60-70 dB). Listeners

were allowed to listen to each voice sample multiple times, if necessary, but only in isolation. That is, listeners could not rate one voice sample, move on to another, and then go back and re-rate the first voice sample. This is to prevent the process of refinement of internal anchors during the rating process. The listeners were blinded to the fact that these voice samples were collected from NTS individuals. The ratings were collected and organized into an excel file for further analysis.

Analysis

Variability

Various descriptive statistical analyses were conducted to further describe the data set and characterize the variability in ratings. Inferential statistical analysis was considered inappropriate for this study because of a lack of homogeneity of variability was hypothesized. Also, due to the relatively small sample size ($n = 6$), the power to detect differences through inferential statistics would be insufficient.

Mean and standard deviations of each of the individual listeners ratings were analyzed. This provided a detailed look at the trends of the individual listeners. Second, this study analyzed the variability of the two groups' ratings. Variability is the extent to which points of data are distant from one another, which is important to understanding consistency and accuracy. This study reported variability in terms of standard deviance, range, interquartile range, and variance. Standard deviation demonstrates the average distance a data point is from the mean data score. Range is a simple and intuitively understood measure of variability. Interquartile range demonstrates the variability of the middle half of the distribution, which contains most of the data points and dismisses outlying data points. Variance reports the spread of the data points within a set.

Reliability

In both the experienced clinicians group and the inexperienced clinicians group, percent agreement was used to calculate reliability. Other methods of calculating reliability were considered, such as ICC, but due to the heavily skewed nature of the data (due to the sample being only NTS individuals), percent agreement was determined to be the most appropriate. Kappa statistics, though useful in determining chance agreement, would be depreciated due to the nature of this study. Because the CAPE-V is a 100-pt scale and because the study focused on voices which (if rated accurately) should result in ratings skewed towards the left end of the scale, the data was necessarily skewed, and correlational statistics would not be effective descriptors of the results.

For the purposes of this study, the acceptable level of reliability was set at 80% agreement. Other studies have used this level of acceptability (Karnell et al., 2007; Kelchner et al, 2010). Clinical rationale also suggests the use of 80% agreement as the acceptable level, as this is strong enough to prevent chance agreement, but also allows for a slight variance in ratings which occur in clinical settings. A moderate level of reliability for this study was set at 60% to 80% agreement. Reliability below 60% was considered an unacceptable level.

Intra-Rater Reliability

The listeners completed a blinded rating of nine randomly selected voice samples that were presented twice in the listening samples, which is 25% redundancy of the total 36 voice samples. In total, then, the listeners rated 45 voice samples. The following parameters were used for analysis: if ratings are 0-9 points different on the 100-pt scale, ratings were treated as being in agreement; if ratings are 10 or more points different, ratings were treated as being dissimilar. This method of differentiation was a hypothetical attempt at finding the balance between the

variability of the inexperienced clinicians group and the experienced clinicians group. If the agreement range were smaller, then an artificially reduced reliability may have been measured. If the agreement range was larger, then an inflated reliability may have been measured. The number of dissimilar ratings was assessed, and percent agreement was then calculated and reported similar to the analysis completed by Bassich and Ludlow (1986).

Inter-Rater Reliability

Each listener's ratings were compared to each other listener's ratings within their respective group. For the experienced clinician group, Experienced Listener A's ratings were compared to those of Experienced Listener B (eAB) and Experienced Listener C (eAC). Experienced Listener B's ratings were then be compared to Experienced Listener C's ratings (eBC). For the inexperienced clinician group, the three pairings were similar to that of the experienced clinician group, and were labeled similarly (iAB, iAC, and iBC). Within all six of those pairings, the following parameters were used for analysis: if ratings are 0-9 points different on the 100-pt scale, ratings were treated as being in agreement; if ratings are 10 or more points different, ratings were treated as being dissimilar. The number of dissimilar ratings were assessed, and percent agreement was calculated and reported, similar to the analysis completed by Bassich and Ludlow (1986).

CHAPTER 3:

RESULTS

Mean Ratings

The mean rating for each voice parameter for each rater are displayed in Figure 1. The inexperienced listeners are represented by lines labeled Inexp. L1, Inexp. L2, and Inexp. L3. The experienced listeners are represented by lines labeled Exp. L1, Exp. L2, and Exp. L3. Though the CAPE-V potentially produces ratings of 0-100 (Kempster et al., 2009), the scale of the y-axis for Figure 1 is 0-30 due to the skewed nature of the data collected. Even though the voice samples rated by the listeners were of NTS individuals, there were mean ratings well above the non-pathological ratings of the CAPE-V, as determined by a clinically significant 10-pt difference. Breathiness appeared to be most consistently reported across the two groups as being non-pathological; roughness and strain appeared to have the greatest variability in the ratings given.

Table 2 contains the mean ratings and the standard deviations for the ratings of each vocal characteristic for each of the listeners. Inspection of these individual listener ratings did not reveal obvious differences due to listener experience.

Figure 1:

Mean ratings by individual listeners for each vocal characteristic

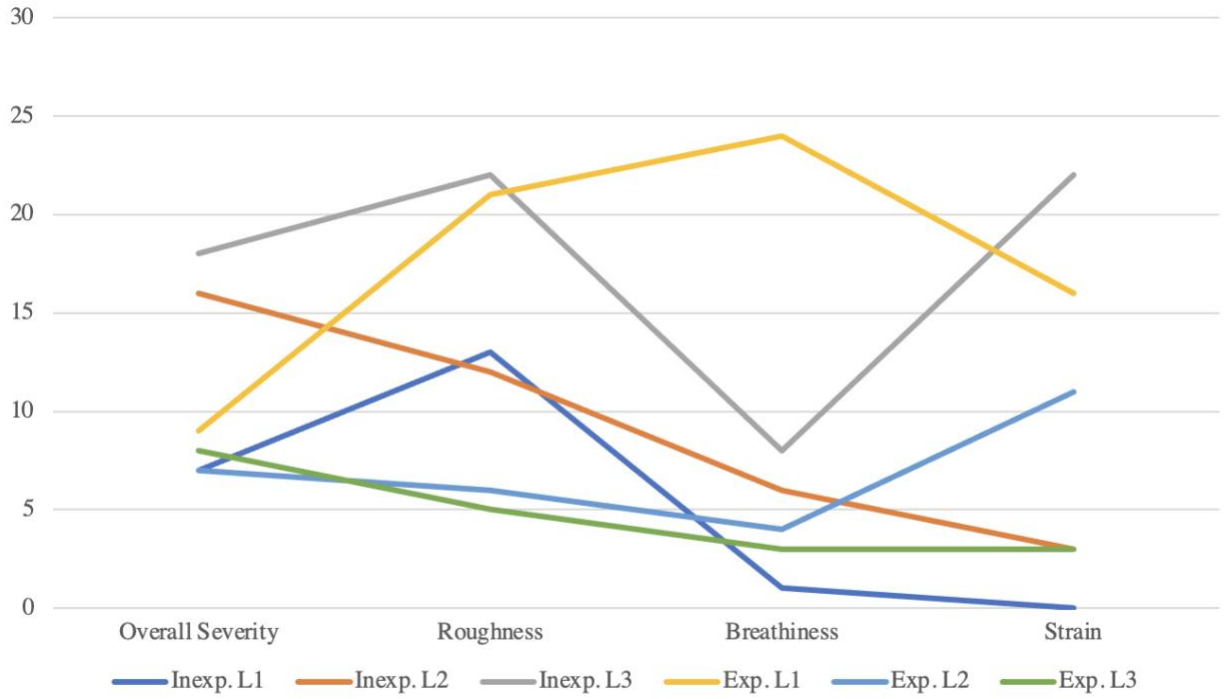


Table 2:

Mean ratings by individual listeners for each vocal characteristic

		Overall Severity	Roughness	Breathiness	Strain
Inexp. Listener 1	Mean	7.29	13.00	1.00	0.00
	SD	7.83	12.04	3.89	0.00
Inexp. Listener 2	Mean	16.47	11.60	6.24	2.73
	SD	13.23	13.14	9.54	5.78
Inexp. Listener 3	Mean	17.56	22.18	7.69	22.13
	SD	11.67	15.74	12.30	15.28
Exp. Listener 1	Mean	8.89	20.73	24.11	16.47
	SD	4.41	9.99	9.37	11.05
Exp. Listener 2	Mean	7.40	5.60	3.80	10.80
	SD	7.13	5.76	6.05	8.45
Exp. Listener 3	Mean	8.18	4.89	2.84	2.76
	SD	5.42	3.96	3.70	3.01

Variability

Variability is used in statistical analysis of data to describe the spread of the data – or how much variation there is within the set. For this study, standard deviation, range, interquartile range, and variance will be reported and discussed. Variability of the mean ratings of the inexperienced clinicians group and the experienced clinicians group are presented in Tables 3 and 4, respectively. Data are reported for each of the vocal characteristics that were rated.

Standard deviation demonstrates how close the data points are to the mean of the data set. A smaller standard deviation denotes a tighter grouping around the mean. As can be seen in Tables 3 and 4, the experienced group had a smaller standard deviation in overall severity, roughness, and strain; meaning the average ratings given were closer to the mean rating for each of those three vocal characteristics. Only in breathiness did the inexperienced clinician group have a smaller standard deviation.

Range is simply the difference between the greatest and least data points within a set. A smaller range indicates less variability. Tables 3 and 4 report both the minimum and maximum data points for each vocal characteristic measured. Because each minimum data point has a value of 0, the maximum data point value also happens to be the value of the range. The data shows that the experienced clinician group had smaller range for each of the four vocal characteristics.

Interquartile range illustrates where the middle 50% of the data points lie in a data set. This is useful to remove the effects of "outliers" (data points that fall in the outer quartiles and thus have significant effects on statistical analysis of the data set). The experienced clinician group had lower interquartile range for overall severity, roughness, and strain; the inexperienced clinician group had a lower interquartile range for breathiness (Tables 3 and 4). This demonstrates the usefulness of removing outliers in data sets – the above-reported range in the

inexperienced clinician group was significantly impacted by these outliers, as can be seen by the greater similarity between their interquartile range and that of the experienced clinician's group.

Variance is a measurement of how spread out the data points are in a data set. The greater the variance, the wider the spread. As shown in Tables 3 and 4, the experienced clinicians group had a smaller variance measurement reported for overall severity, roughness, and strain; the inexperienced clinician's group had a smaller variance reported for breathiness.

Table 3:

Variability of mean ratings of inexperienced clinician group for each vocal characteristic

	Overall Severity	Roughness	Breathiness	Strain
Mean	13.77	15.59	4.98	8.29
SD	12.06	14.51	9.70	14.19
Range	*Min* 0 *Max* 57	*Min* 0 *Max* 67	*Min* 0 *Max* 45	*Min* 0 *Max* 66
Interquartile range	15.00	20.00	6.00	14.00
Variance	145.38	210.46	94.07	201.32

Table 4:

Variability of mean ratings of experienced clinician group for each vocal characteristic

	Overall Severity	Roughness	Breathiness	Strain
Mean	8.17	10.41	10.25	10.01
SD	5.80	10.15	11.93	9.96
Range	*Min* 0 *Max* 31	*Min* 0 *Max* 42	*Min* 0 *Max* 41	*Min* 0 *Max* 38
Interquartile range	7.00	15.00	20.00	13.00
Variance	33.58	102.99	142.22	99.10

Reliability

Intra-Rater Reliability

The percent agreement values for intra-rater reliability are reported in Table 5. The experienced clinicians group had higher intra-rater reliability in all four vocal characteristics when compared to the inexperienced clinicians group as reflected in the percent agreement results. The magnitude of difference in the percent agreement between the experienced clinicians group and the inexperienced clinicians group was 41% (overall severity), 25% (roughness), 26% (breathiness), and 8% (strain). Across vocal characteristics, the experienced group had percent agreement values $\geq$ 80%. Ratings of overall severity had the highest percent agreement in the experienced group and strain and roughness had the lowest. In contrast, percent agreement values ranged from 59% to 74% for the inexperienced group with the highest value for strain and the lowest for overall severity and roughness.

Table 5:

Intra-rater reliability of ratings for each vocal characteristic and group reported as percent agreement

	Inexperienced Clinicians Group	Experienced Clinicians Group
Overall Severity	59%	100%
Roughness	59%	85%
Breathiness	67%	93%
Strain	74%	82%

Inter-Rater Reliability

Percent agreement values for inter-rater reliability analysis are reported in Table 6. Inter-rater reliability was higher for the experienced clinician group for overall severity (31% higher), roughness (9% higher), and strain (5% higher). The inexperienced clinician group had higher percent agreement for breathiness (29% higher). The experienced clinicians group had greater

inter-rater reliability for rating overall severity, roughness, and strain when compared to the inexperienced clinicians group. The inter-rater reliability for rating breathiness was greater in the inexperienced clinicians group. The experienced group had percent agreement values ≥ 80% for only overall severity; percent agreement for roughness, breathiness, and strain were <60%. Ratings of overall severity had the highest percent agreement in the experienced group and roughness and breathiness had the lowest. In the inexperienced group, percent agreement values for breathiness were >60%, though less than 80%. Percent agreement for overall severity, roughness, and strain were all <60%. Ratings for breathiness had the highest percent agreement in the inexperienced group and overall severity and roughness had the lowest.

Table 6:

Inter-rater reliability of ratings for each vocal characteristic and group reported as percent agreement

	Inexperienced Clinicians Group	Experienced Clinicians Group
Overall Severity	50%	81%
Roughness	39%	48%
Breathiness	69%	40%
Strain	47%	52%

As has been discussed, inferential statistics would be an ineffective way to analyze this data set. To provide further confirmation of this, calculations for kappa statistics can be found in Appendix E.

CHAPTER 4:

DISCUSSION

Auditory-perceptual voice assessments are often used by speech-language pathologists in assessing the voice and for planning and adjusting treatment goals (Bassich & Ludlow, 1986; Yamasaki et al., 2017). Further, this method is considered to be the gold standard for assessing the voice by some (Chan & Yiu, 2002; Helou et al., 2010). The CAPE-V is an auditory-perceptual voice assessment that has grown in popularity within clinical practice, in part because of the care taken in constructing the tool and assessing its psychometric properties (Kempster et al., 2009; Helou et al., 2010). As such, research on the reliability of these assessments has implications on the well-being of clients and on the efficacy of speech-language pathologists as clinicians. Because these assessments are used (at times solely) to develop treatment goals, accurate assessments of a client's needs are critical to achieve their goals. Clients with both non-pathological and mildly dysphonic voices would benefit from reliable assessments because clinicians theoretically would be prevented from suggesting treatment which is not truly needed.

This study sought to fill a void in current literature by studying the variability and reliability of the CAPE-V when it is used to assess non-pathological voices as judged by experienced and inexperienced clinicians. These clinicians performed ratings on 45 voice samples of NTS individuals using the CAPE-V, providing ratings for overall severity, roughness, breathiness, and strain.

The research questions centered around the variability and reliability of the CAPE-V with the dependent variable being the experience of the clinician. For the research question concerning variability (RQ1), it was hypothesized that the inexperienced clinicians group will demonstrate greater variability in all four vocal characteristics studied when compared to the

experienced clinicians group. This research question was addressed by studying the variability of the data – specifically the standard deviation, range, interquartile range, and variance. This was based on the literature review, where one study was conducted in a similar manner (Helou et al., 2010).

For the research question concerning reliability (RQ2), it was hypothesized that that intra- and inter-rater reliability will be greater in the experienced clinicians group. That is, the inexperienced clinicians will demonstrate less agreement in the ratings of non-pathological voices across all four vocal characteristics rated. This research question was tested by calculating the percent agreement to analyze the intra- and inter-reliability of the two groups. This was based on the review of the literature, where one study found similar results when using pathological voice samples (Helou et al., 2010).

Summary of Main Findings

Research Question 1 (Variability)

H1 (concerning variability in rating overall severity) was supported by the descriptive statistical analysis of the data set. All four measures of variability (standard deviation, range, interquartile range, and variance) suggested that the inexperienced clinicians demonstrate greater variability when rating overall severity in non-pathological samples with the CAPE-V.

H2 (concerning variability in rating roughness) was supported by the descriptive statistical analysis of the data set. All four measures of variability suggested that the inexperienced clinicians demonstrate greater variability when rating roughness in non-pathological samples with the CAPE-V.

H3 (concerning variability in rating breathiness) was not supported by the descriptive statistical analysis of the data set. Only the range suggested that the inexperienced clinicians demonstrate greater variability when rating breathiness in non-pathological samples with the CAPE-V. The remaining three measures of variability suggest that the experienced clinicians display more variability.

H4 (concerning variability in rating strain) was supported by the descriptive statistical analysis of the data set. All four measures of variability suggested that the inexperienced clinicians demonstrate greater variability when rating strain in non-pathological samples with the CAPE-V.

Research Question 2 (Reliability)

H5 (concerning reliability in rating overall severity) was supported by the percent agreement values calculated for both intra- and inter-rater reliability. Percent agreement for the intra-rater reliability of the inexperienced and experienced clinicians was 59% and 100%, respectively. The level of intra-rater reliability for the inexperienced clinicians fell below the acceptable level, but for the experienced clinicians it was considered acceptable. Percent agreement for the inter-rater reliability of the inexperienced and experienced clinicians was 50% and 81%, respectively. The level of inter-rater reliability for the inexperienced clinicians fell below the acceptable level, but for the experienced clinicians it was considered acceptable.

H6 (concerning variability in rating roughness) was supported by the percent agreement values calculated for both intra- and inter-rater reliability. Percent agreement for the intra-rater reliability of the inexperienced and the experienced clinicians was 59% and 85%, respectively. The level of intra-rater reliability for the inexperienced clinicians fell below the acceptable level, but for the experienced clinicians it was considered acceptable. Percent agreement for the inter-

rater reliability of the inexperienced and the experienced clinicians was 39% and 48%, respectively. The level of inter-rater reliability for the inexperienced and experienced clinicians both fell below the acceptable level.

H7 (concerning variability in rating breathiness) was partially supported. Percent agreement for the intra-rater reliability of the inexperienced and the experienced clinicians was 67% and 93%, respectively. These values supported H7. The level of intra-rater reliability for the inexperienced clinicians were moderately acceptable, and for the experienced clinicians it was considered acceptable. Percent agreement for the inter-rater reliability of the inexperienced and experienced clinicians was 69% and 40%, respectively. These values did not support H7. The level of inter-rater reliability for the inexperienced clinicians was moderately acceptable, but for the experienced clinicians it was considered unacceptable.

H8 (concerning variability in rating strain) was partially supported. Percent agreement for the intra-rater reliability of the inexperienced and the experienced clinicians was 74% and 82%, respectively. These values supported H8. The level of intra-rater reliability for the inexperienced and experienced clinicians were moderately acceptable. Percent agreement for the inter-rater reliability of the inexperienced and the experienced clinicians was 47% and 52%, respectively. These values did not support H8. The level of inter-rater reliability for the inexperienced and experienced clinicians was considered unacceptable.

Further Discussion of Main Findings

Table 7 provides a summary of the findings detailed previously. It lists each individual hypothesis and whether that hypothesis is supported by its applicable measurement (i.e., variability, or intra- and inter-rater reliability). Table 7 then clearly whether each hypothesis was fully supported (F), partially supported (P), or not supported (N).

Table 7:

Summary of measurement findings and the support of the hypotheses

		Measurement contributed to supporting hypothesis?			**Hypothesis supported?**
		Variability	Intra-Rater Reliability	Inter-Rater Reliability	
RQ1	H1	*Yes*	*n/a*	*n/a*	*F*
	H2	*Yes*	*n/a*	*n/a*	*F*
	H3	*Yes*	*n/a*	*n/a*	*N*
	H4	*Yes*	*n/a*	*n/a*	*F*
RQ2	H5	*n/a*	*Yes*	*Yes*	*F*
	H6	*n/a*	*Yes*	*Yes*	*F*
	H7	*n/a*	*Yes*	*No*	*P*
	H8	*n/a*	*Yes*	*No*	*P*

These findings reflect those of similar studies. In Helou et al. (2010), the inexperienced clinicians also demonstrated low inter-rater reliability. However, Helou et al. (2010), Kelchner et al. (2010), Kempster et al. (2009), and Nemr et al. (2012) found strong inter-rater reliability in the ratings of experienced clinicians. In this study, only ratings of overall severity had an acceptable (i.e., strong) level of reliability in the experienced clinicians group's data. This disparity could be a result of the small sample size of this study.

It is interesting to note that, in this study, only overall severity enjoyed an acceptable level of both intra- and inter-rater reliability when the ratings are completed by the experienced clinicians. All the other vocal characteristics ratings do not have strong levels of reliability, whether they are completed by the inexperienced or the experienced clinicians. This is perhaps suggestive of the fact that the assessment of the deviance of the voice as a whole is easier, because it takes into account all of the individual vocal characteristics (e.g., roughness, breathiness, and strain). This is not the only study with these findings; Eadie and Baylor (2006) and Helou et al. (2010) reported parallel reliability levels for the aforementioned vocal characteristics. There may be a way to capitalize on this finding, such as developing an auditory-

perceptual assessment that focuses on the overall quality of the voice, instead of individual vocal characteristics. If strong intra- and inter-reliability can be attained and maintained, both clinicians and clients will benefit.

Another interesting observation is the potential effects of gender and age on the ratings. Although the targeted variable was experience, another variable that could have potentially been studied was the gender of the clinicians. In the inexperienced clinician group, there were two females and one male; in the experienced clinician group, all three listeners were male. Looking at Figure 1 on page 20, it is hard to discern any distinct differences between the listeners based on gender alone. Also, the distinction between the female voice samples (which comprised 52.8% of the voice sample population) and the male voices (47.2%) may have an effect on the data. Research has shown that certain vocal characteristics are more culturally acceptable in one gender than the other. For example, Van Borsel et al. (2009) found that breathiness is a component of the femininity of a speaker. In this present study, then, the listeners may be more apt to rate a voice as less deviant in breathiness because the voice is feminine, and it is culturally acceptable for a female voice to be breathier. It is difficult to know this for sure. Further analysis would need to be done on this variable, as gender was not the focus of this present study.

Age could also have had an influence on the ratings. As is discussed later in this chapter, research has shown that age affects the ability to perceive auditory stimuli (Goy & Pichora-Fuller, 2016). Examiinng Figure 1 on page 20 would reveal perhaps one example of the effects of age on rating breathiness. Exp. Listener 1 (EL1) was older than not only the three inexperienced listeners, but also the other two experienced clinicians. Also, breathiness is the vocal characteristic which EL1's ratings were significantly distant from those of the other two experienced listeners. Breathiness is a vocal characteristic which requires the ability to hear high

frequency sounds. As an individual ages, they may lose the ability to perceive those high frequency sounds. These two factors may have played a role in the data set not supporting H3. However, one would expect the ratings to be significantly lower than the mean if EL1 could not perceive the high frequency sounds of breathiness. This, though, is not the case, as EL1's average rating was13.86 points above the mean average rating for breathiness. Perhaps overcompensation of the inability to accurately perceive auditory stimuli played a role in this strong trend towards higher ratings. Further analysis would need to be done on this variable, as age was not the focus of this present study.

Limitations of Present Study

The most significant limitation encountered in the conduction of this study is the small sample size of the listeners. With both groups only consisting of three clinicians each, inferential statistics were unable to be employed. This made it difficult to draw conclusions from the data collected.

A second limitation encountered was that of the identification of the voice samples as non-pathological. No method of ensuring that these voices were clinically non-pathological was employed, leaving that qualification to be met by the self-perception of the voice sample participant. As a result, the participants could only be identified by this study as being "non-therapy seeking", so the potential exists for a pathological voice (albeit undetected or considered benign by the participant) to be in the voice sample collection.

A third limitation is the lack of a control group, which would be a set of moderately to severely pathological voice samples. Having the two groups of clinicians rate pathological voices would have resulted in intra- and inter-reliability measures to which this study would have compared the intra- and inter-reliability measures of the ratings of the non-pathological voices.

This would have allowed further analysis to be done, isolating the presence of pathology as a factor in the reliability of the CAPE-V.

A fourth and final limitation of this study is the heavily skewed nature of the data. Results of skewness analysis have been included in Appendix F to further confirm this limitation Due to the nature of this study, this is an inescapable limitation – it is an outcome of the research question being investigated. For the voice samples to be accurately rated, only a portion of the CAPE-V would be used. This, however, resulted in an inability to utilize inferential statistics in the analysis of the data.

Directions for Future Research

Because this study used the CAPE-V, an auditory-perceptual voice assessment typically used to assess pathological voices, to assess non-pathological voice samples, and due to the fact that the listeners were blinded to this fact, it would be beneficial if future research investigated the innate characteristics of listeners. Both Warren (1967) and Kent (1996) write that psychology – expectations, presuppositions, and predispositions – play an important role in perceptual measures. A post-experimental survey sent to the listeners to assess their expectations going into the rating tasks, and to assess their assumptions after the fact, may prove beneficial in understanding the decreased reliability measured in this current study. Examples of questions which could be included in the survey are: “would you classify this voice (or these voices) as pathological? If so, how severe?” and “to what extent did your expectations of these voice samples change as you continued through this study and at its conclusion?”. These types of questions would suggest the preconceived ideas that the listeners had when beginning the rating tasks, and if (and how) those preconceptions changed as a result of the study.

Another direction for future research would be to study the effects of age on the use of auditory-perceptual voice assessments. A study conducted in a field peripheral to speech-language pathology found that older clinicians rated stimuli differently than younger clinicians (Goy & Pichora-Fuller, 2016). One issue that could be faced is the limited number of younger clinicians with large amounts of experience. This problem can be mitigated by providing extensive and intensive training to newly graduated clinicians – perhaps even up to a year. Alternatively, the sample could include only younger clinicians who have an unusually large amount of experience in the field of voice.

Conclusion

This study focused on the use of the CAPE-V on non-pathological voices and how experience of the clinician administering the assessment affected the ratings. The findings suggest that the more experience that a clinician has, the more reliable the ratings will be – as measured by variability and intra-reliability in this study – at least for the vocal characteristics of overall severity, roughness, and strain. However, the findings in this present study do not suggest strong inter-rater reliability, contrary to the findings of Helou et al. (2010), Kelchner et al. (2010), Kempster et al. (2009), and Nemr et al. (2012). This is possibly due to a small sample size. More research should be conducted on the use of CAPE-V – and other auditory-perceptual voice assessments – on clients with non-pathological and mildly disordered voices.

APPENDICES

APPENDIX A:

GRBAS PROTOCOL

GRBAS

Component	0 Normal	1 Slight	2 Moderate	3 Extreme
G Overall Grade of hoarseness				
R Rough				
B Breathy				
A Aesthenic				
S Strained				

Procedures: Grade the five voice parameters using the four point scale. *Place a check mark in the column indicating the perceived severity of each parameter.*

G: represents the degree of hoarseness or voice abnormality

R: represents a psycho-acoustic impression of the irregularity of vocal fold vibrations. It corresponds to the irregular fluctuations in the fundamental frequency and/or amplitude of the glottal sound source.

B: represents a psycho-acoustic impression of the extent of air leakage through the glottis. It is related to turbulence.

A: denotes weakness or lack of power in the voice. It is related to a weak intensity of the glottal source sound and/or lack of higher harmonics.

S: represents a psycho-acoustic impression of a hyperfunctional state of phonation. It is related to an abnormally high fundamental frequency, noise in the high frequency range, and/or richness in high frequency harmonics.

Score: A voice may be judged to be "1" for grade (i.e., slight), "1" for rough, "1" for breathy, "0" for aesthenic, and "0" for strained (G1R1B1A0S0) or any combination of the above.

Hirano, 1981

APPENDIX B:

BUFFALO III VOICE PROFILE PROTOCOL

Voice Rating Scale

Student: _____ Date: _____
DOB: _____ CA: _____ Grade: _____
Teacher: _____ School: _____

Circle the score for each of the five parameters listed. Add the five scores to determine the total score and total voice rating.

BUFFALO III VOICE PROFILE **Pitch**	**Score 0** Pitch is within normal limits **BIIIVP Rating 1**	**Score 1** Pitch is noticeably different, but intermittent. Pitch is not considered distracting or an interference to communication **BIIIVP Rating 2**	**Score 2** Pitch is persistently different (too high or low) and inappropriate to age and gender and interferes with communication **BIIIVP Rating 3**	**Score 3** Pitch is persistently different and inappropriate to age and gender and greatly interferes with communication **BIIIVP Rating 4-5**
BUFFALO III VOICE PROFILE **Intensity**	**Score 0** Intensity is within normal limits **BIIIVP Rating 1**	**Score 1** Intensity is noticeably different, but intermittent. Intensity is not considered distracting or an interference to communication **BIIIVP Rating 2**	**Score 2** Intensity is persistently too loud, too soft, or dysphonic; inappropriate to situations and interferes with communication **BIIIVP Rating 3**	**Score 3** Intensity is persistently too loud, soft, or dysphonic; inappropriate to situations and greatly interferes with communication **BIIIVP Rating 4-5**
BUFFALO III VOICE PROFILE **Quality**	**Score 0** Quality is within normal limits **BIIIVP Rating 1**	**Score 1** Quality is noticeably different, but intermittent. Quality is not considered distracting or an interference to communication **BIIIVP Rating 2**	**Score 2** Quality is persistently hoarse, breathy, tense, strident or contains other abnormal attributes; inappropriate for age and gender; interferes with communication **BIIIVP Rating 3**	**Score 3** Quality is persistently hoarse, breathy, tense, strident, or contains other abnormal attributes; inappropriate for age and gender; greatly interferes with communication **BIIIVP Rating 4-5**
BUFFALO III VOICE PROFILE **Resonance**	**Score 0** Resonance is within normal limits **BIIIVP Rating 1**	**Score 1** Resonance is noticeably different, but intermittent. Resonance is not considered distracting or an interference to communication **BIIIVP Rating 2**	**Score 2** Resonance is persistently different and inappropriate; interferes with communication **BIIIVP Rating 3**	**Score 3** Resonance is persistently different and inappropriate; greatly interferes with communication **BIIIVP Rating 4-5**
TEACHER INPUT FORM FOR VOICE Adverse Affect on Educational Performance Social -Emotional Academic -Vocational	**Score 0** No interference with student's participation in educational settings	**Score 2** Minimal impact on student's participation in educational settings	**Score 3** Interferes with student's participation in educational settings	**Score 4** Greatly interferes with student's participation in educational settings
TOTAL SCORE	**0 1 2 3 4**	**5 6 7 8**	**9 10 11 12**	**13 14 15 16**
TOTAL VOICE RATING	**Non-handicapping**	**Mild**	**Moderate**	**Severe**

The student:
- ❑ **meets** the eligibility criteria for speech and language services in the area of voice.
- ❑ **does not meet** the eligibility criteria for speech and language services in the area of voice.

Comments: _____

Wilson, 1987

APPENDIX C:

CAPE-V PROTOCOL

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:_____

Kempster et al., 2009

APPENDIX D:

RAINBOW PASSAGE

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Others have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbows. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of super-imposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

Fairbanks, 1960

APPENDIX E:

INTER-RATER RELIABILITY CALCULATED USING FLEISS' KAPPA

Another method of analyzing inter-rater reliability is through the use of kappa statistics. Due to the skewed nature of this data and the limited sample size, this method of data analysis failed to illustrate anything noteworthy. However, to provide resources for further investigation, the following data has been included as an appendix.

In this case, because there were more than two listeners for whom to analyze inter-rater reliability, Fleiss' kappa is best suited for this task (Fleiss, 1971). Table 8 displays the kappa coefficients and the significance (as p-values) of both groups inter-reliability, separated by the four vocal characteristics. Given that negative kappa coefficients indicate that agreement between listeners is less than agreement expected by chance (Fleiss, 1971), only the ratings for overall severity given by the experienced clinician group are not attributable to chance. *P*-values were reported to show statistical significance (where $p < 0.05$). As such, with the exception of the ratings given for roughness and breathiness by the inexperienced clinicians group, none of the Fleiss' kappa coefficients are statistically significant.

Table 8:

Inter-rater reliability (Fleiss' kappa)

	Inexperienced Clinicians Group		Experienced Clinicians Group	
	Kappa	*P*-value	Kappa	*P*-value
Overall Severity	-0.001	0.971	0.027	0.229
Roughness	-0.001	0.005	-0.020	0.319
Breathiness	-0.088	0.007	-0.013	0.635
Strain	-0.104	0.960	-0.025	0.248

APPENDIX F:

SKEWNESS GRAPHS BY VOCAL CHARACTERISTIC AND EXPERIENCE

Skewness measures the asymmetrical distribution of a data set. In a skewed data set, statistical analysis that relies on normal distribution will be rendered ineffectual. Most inferential statistical analyses rely on normal distribution and because the majority of the data in this present study is highly right-skewed (as is demonstrated in Table 9), reporting inferential statistical analyses would be meaningless.

Data with skewness of >1 is considered highly skewed. Except for the skewness of the ratings for breathiness given by the experienced clinicians group, all ratings are highly skewed to the right. With a skewness of 0.94, the breathiness ratings given by the experienced listeners is on the higher end of the moderate skewness category.

Table 9:

Skewness statistics

	Overall Severity	Roughness	Breathiness	Strain
Inexp. Listeners	1.18	1.07	2.33	1.93
Exp. Listeners	1.46	1.25	0.94	1.09

Included in this appendix are histograms with overlaid curves to show frequencies and distribution. These charts are representative of data collected on a 100-point scale. However, due to space limitations, the right tail of these charts is cut off, as there is no data to report. In that sense, then, this data is skewed, as is represented in the following charts.

Figure 2:

Skewness of overall severity ratings given by the inexperienced clinicians group

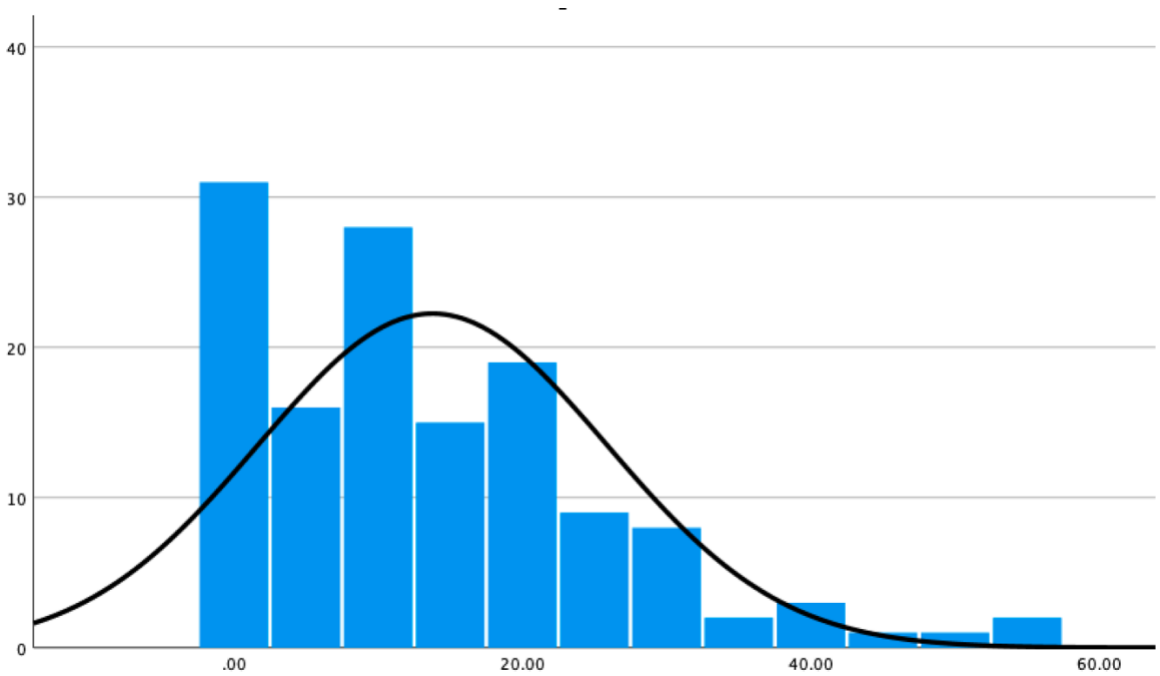


Figure 3:

Skewness of roughness ratings given by the inexperienced clinicians group

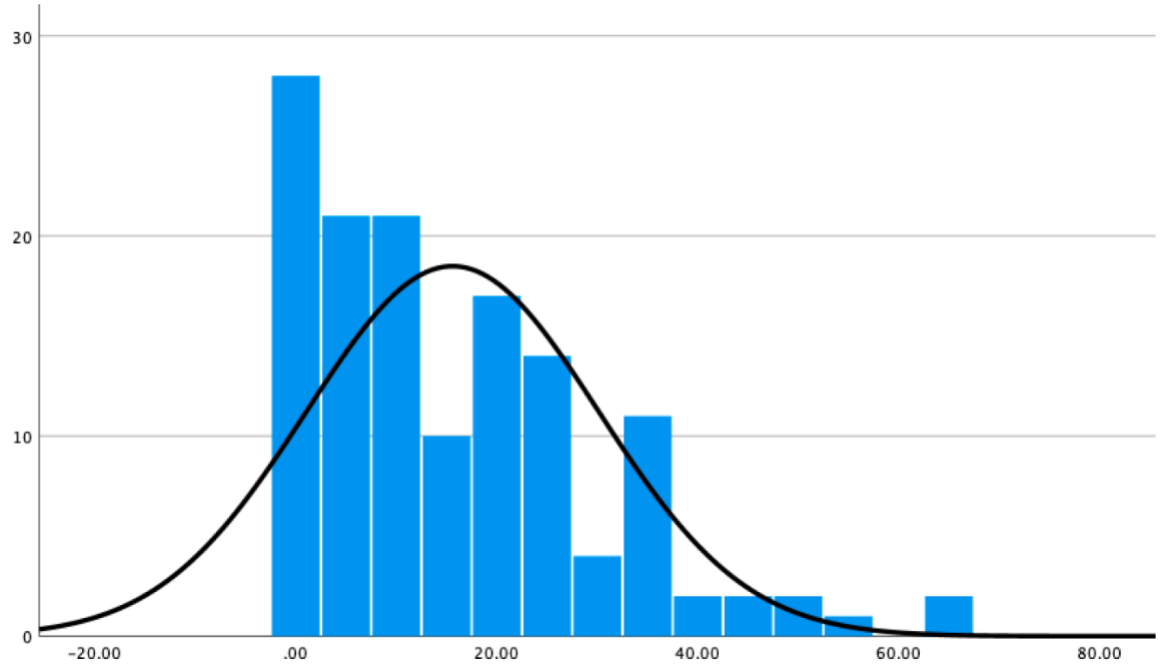


Figure 4:

Skewness of breathiness ratings given by the inexperienced clinicians group

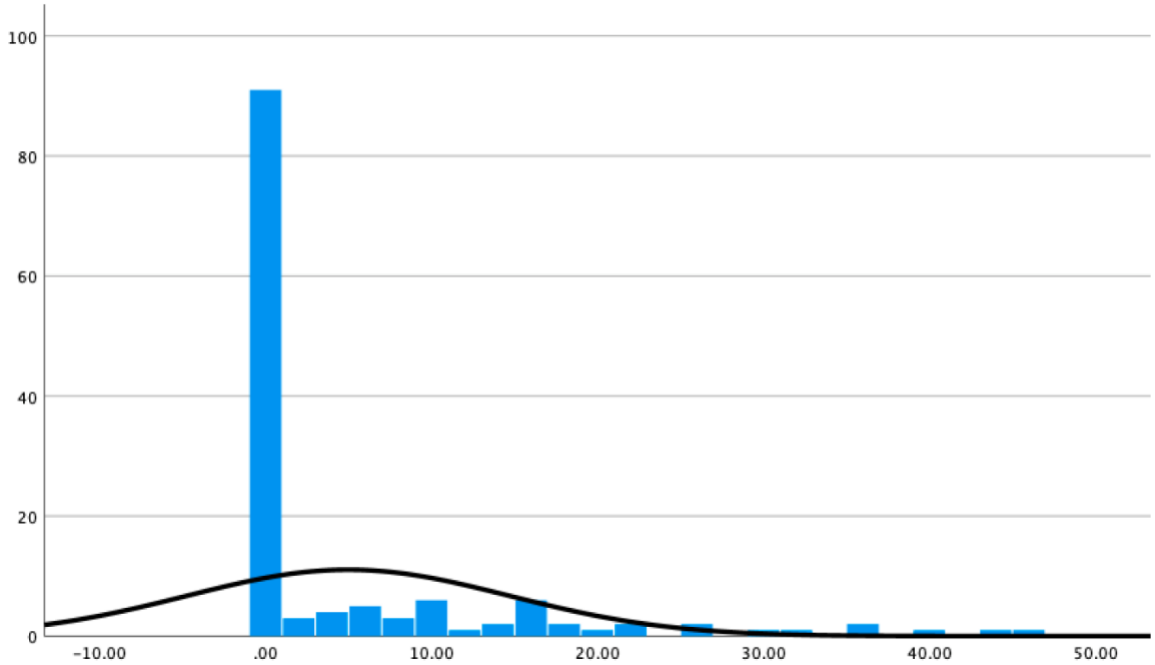


Figure 5:

Skewness of strain ratings given by the inexperienced clinicians group

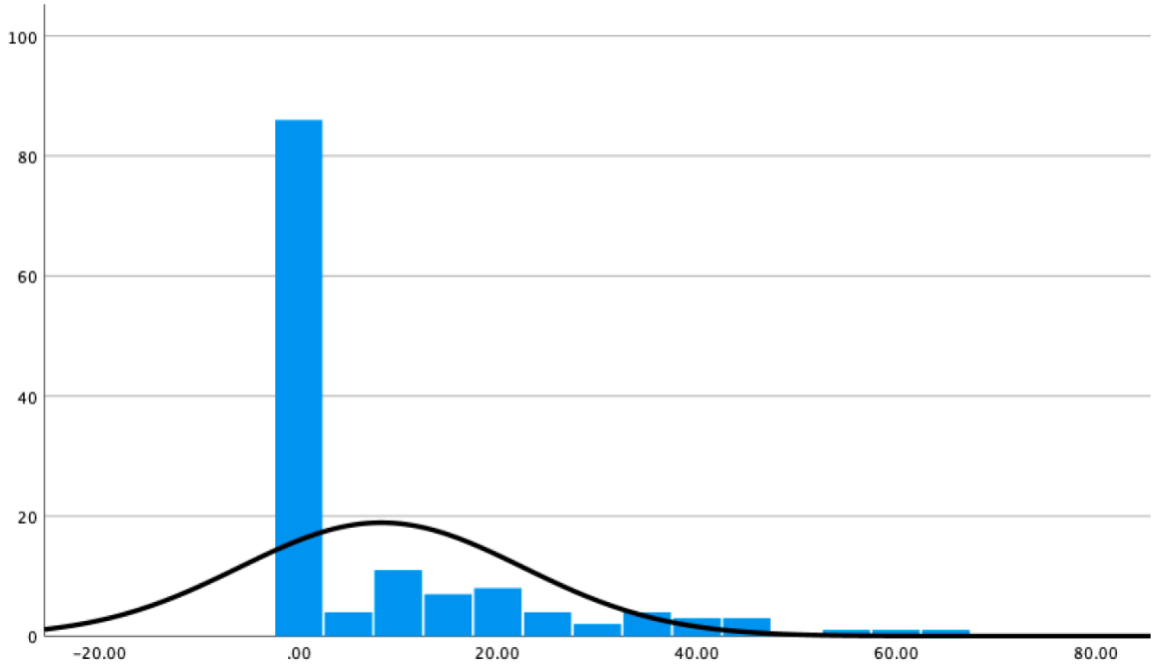


Figure 6:

Skewness of overall severity ratings given by the experienced clinicians group

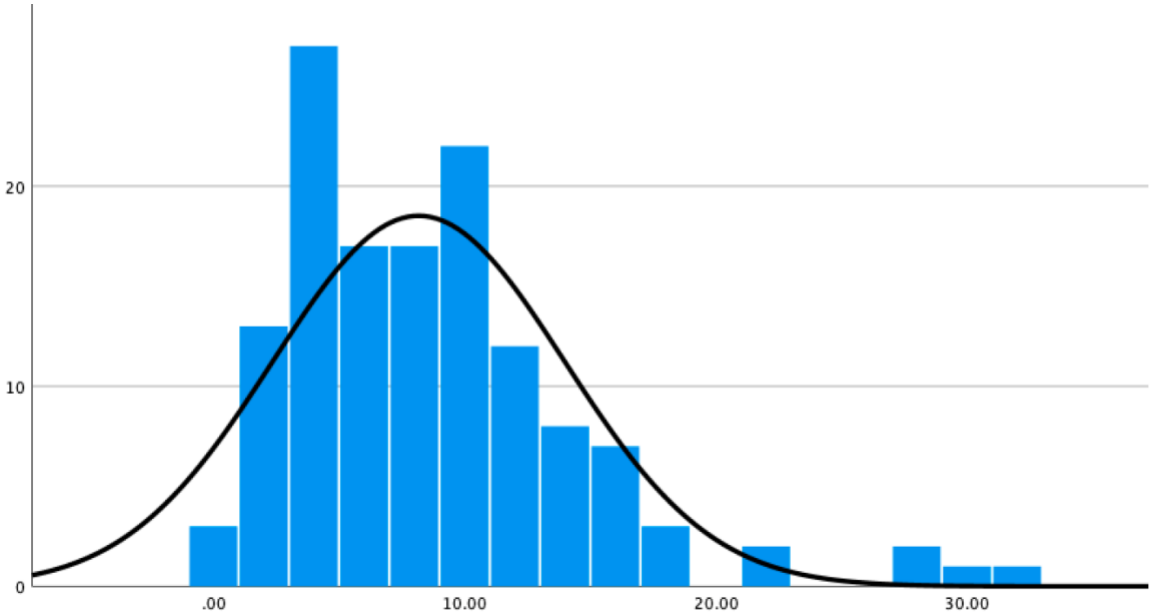


Figure 7:

Skewness of roughness ratings given by the experienced clinicians group

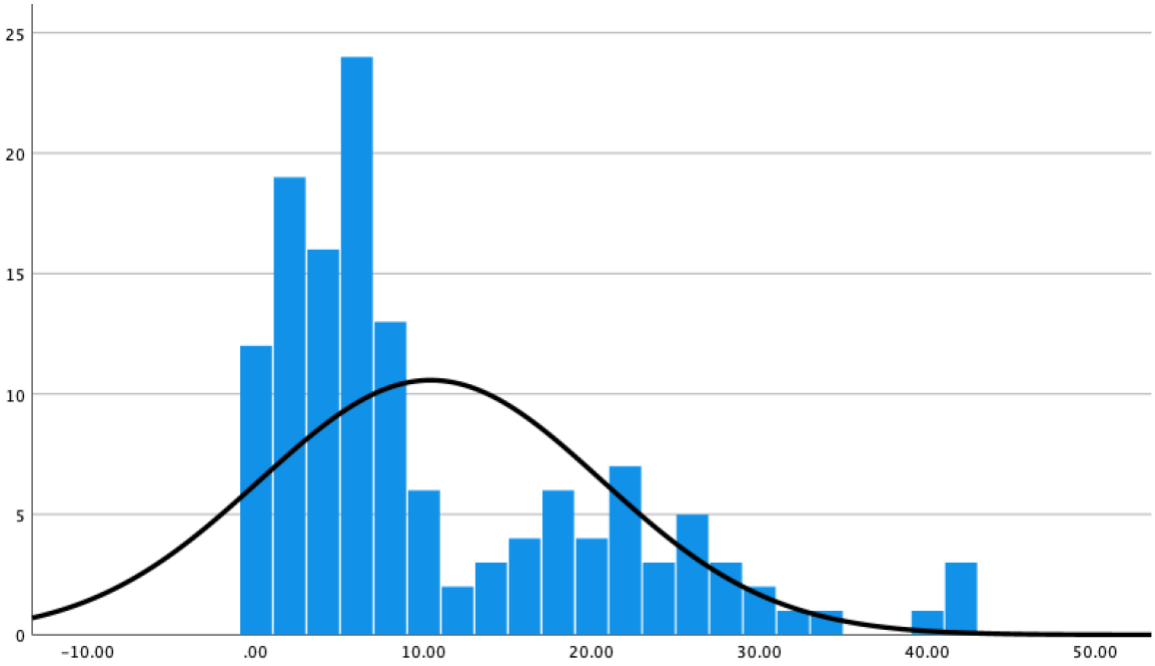


Figure 8:

Skewness of breathiness ratings given by the experienced clinicians group

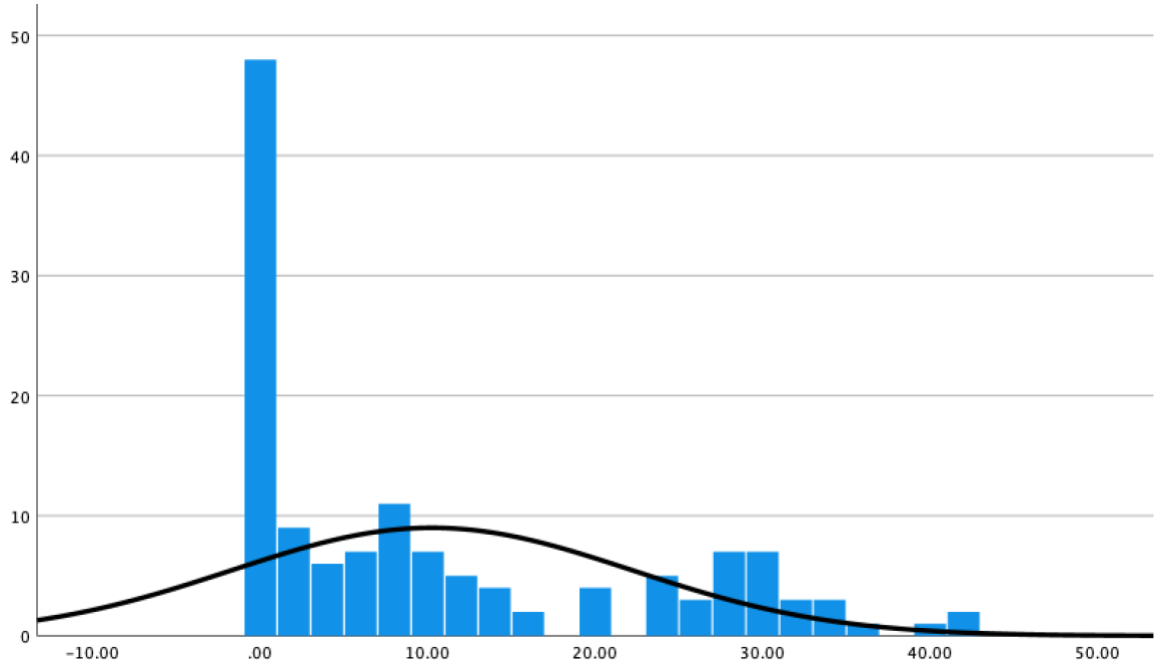
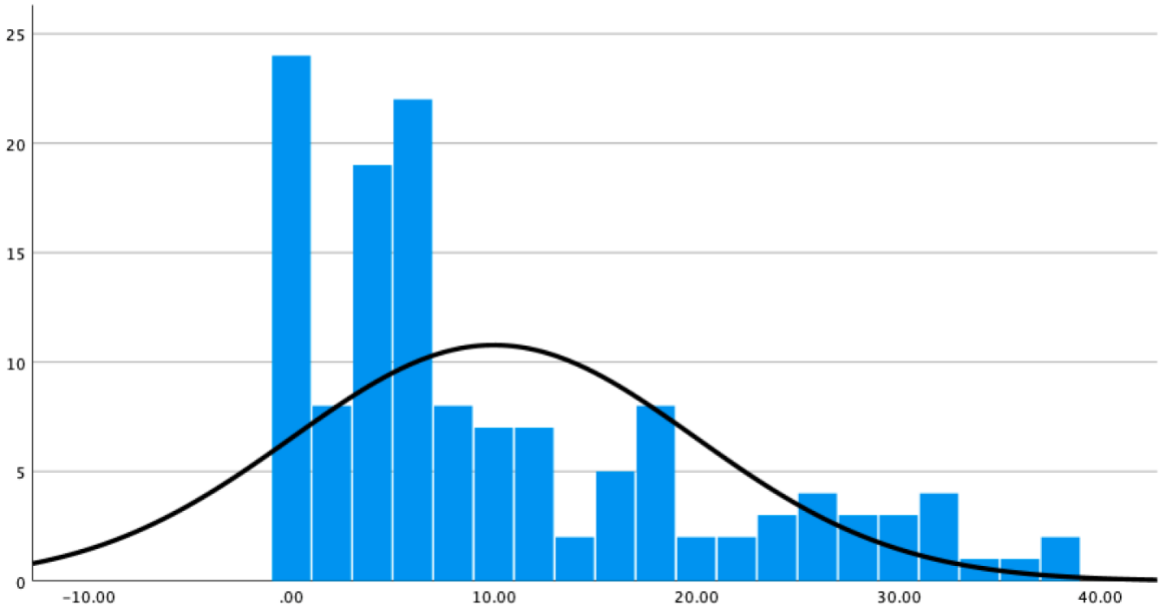


Figure 9:

Skewness of strain ratings given by the experienced clinicians group



REFERENCES

REFERENCES

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.

Bassich, C. J. & Ludlow, C. L. (1986). The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders, 51*, 125-133.

Bele, I. V. (2005). Reliability in perceptual analysis of voice quality. *Journal of Voice, 19*(4), 555-573.

Behrman, A. (2005). Common practices of voice therapists in the evaluation of patients. *Journal of Voice, 19*, 454-469.

Chan, K. M. & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45*(1), 111-126.

De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Croux, C. (1997). Test-retest study of the GRBAS Scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11*(1), 74-80.

Eadie, T. L. & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20*(4), 527-544.

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. & Kaspner-Smith, M. (2011). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research, 54*, 430-447.

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). New York, NY: Harper & Row.

Fex, S. (1992). Perceptual evaluation. *Journal of Voice, 6*(2), 155-158.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Ford, D. S. (2021). *Immediate effects of semi-occluded vocal tract exercises and the implications for clinical practice* [Unpublished doctoral dissertation]. Michigan State University.

Goy, H. & Pichora-Fuller, M. K. (2016). Effects of age on speech and voice quality ratings. *The Journal of the Acoustical Society of America, 139*, 1648-1659.

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 the voice*. New York, NY: Springer-Verlag.

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

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 the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice, 24*(4), 441-449.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., 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.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7-23.

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*, 21-40.

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), 17-22.

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.

Van Borsel, J., Janssens, J., & De Bodt, M. (2009). Breathiness as a feminine voice characteristic: A perceptual approach. *Journal of Voice, 23*(3), 291-294.

Warren, R. (1976). Auditory illusions and perceptual processes. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 389-418). New York: Academic Press.

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. *Laryngology, 267*, 429-434.

Wilson, D. K. (1987). *Voice problems of children* (3rd ed.). Baltimore, MD: Williams and Wilkins.

Yamasaki, R., Madazio, G., Leão, S. H. S., Padovani, M., Azevedo, R., & Behlau, M. (2017). Auditory-perceptual evaluation of normal and dysphonic voices using the Voice Deviation Scale. *Journal of Voice, 31*(1), 67-71.

Yiu, E. M.-L., Chan, K. M. K., & Mok, R. S.-M. (2007). Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical Linguistics and Phonetics, 21*(2), 129-145.

Yu, P., Revis, J., Wuyts, F. L., Zanaret, M., & Giovanni, A. (2002). Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatrica et Logopaedica*, *54*, 271-281.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology, 20*, 14-22.