2022

Perceptual bias in ratings of voice quality: sequential effects

BOSTON UNIVERSITY

SARGENT COLLEGE OF HEALTH AND REHABILITATION SCIENCES

Thesis

PERCEPTUAL BIAS IN RATINGS OF VOICE QUALITY: SEQUENTIAL EFFECTS

by

SHANNON LEE

B.A., University of California, Santa Barbara, 2020

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2022

Approved by

First Reader

Cara E. Stepp, Ph.D.
Professor of Speech, Language, and Hearing Sciences

Second Reader

Gerald D. Kidd Jr., Ph.D.
Professor of Speech, Language, and Hearing Sciences

Third Reader

Daniel Buckley, M.S., CCC-SLP
Research Fellow

Fourth Reader

Kimberly Dahl, M.S., CCC-SLP
Doctoral Student

This thesis is dedicated to my parents, Mike & Leslie, for all their love and support, and to my partner, Fred, for sticking with me through it all.

ACKNOWLEDGEMENTS

I would like to thank everyone in my committee for their knowledge and thoughtful feedback. I would also like to thank everyone in the STEPP lab for their support and kind words, especially Kim and Cara without whom I would have been entirely lost. Thank you also to all the SLPs who took time from their busy schedules to participate. Special thanks to my cat, Scrappy, for being my work from home companion.

PERCEPTUAL BIAS IN RATINGS OF VOICE QUALITY:

SEQUENTIAL EFFECTS

SHANNON LEE

ABSTRACT

The purpose of this research project was to determine the existence and type of sequential effect evident in ratings of voice quality by SLPs with and without prior experience rating voices using the CAPE-V overall severity rating. It was expected that SLPs with experience rating voices will demonstrate a small assimilation bias and those without experience rating voices will demonstrate a larger contrast bias. A between-groups method was used where each participant rated moderately dysphonic voices in a neutral, severe, and mild condition. The severe and mild conditions were composed of 5 severely or mildly dysphonic voices after which the moderate voice was rated again to assess any change in rating by condition. All participants demonstrated a contrast bias in the severe condition only. Raters without experience rating voices had a larger contrast bias, as well as those with greater confidence using the CAPE-V tool. The results of this study suggest caution in using a clustered presentation of voices for research purposes as the contrast bias observed would impact the validity of those observations. Knowledge of this effect should be incorporated into future studies of voice ratings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The idea of perceptual biases is an intriguing one. We would all like to trust our brains to interpret information accurately, but in truth our perception can be influenced by multiple factors. One of these is context, which leads to a type of perception bias called sequential effects and is the topic of exploration for this thesis project. The goal is to establish if this type of effect exists for ratings of voice quality, and to what extent. It also hopes to explore potential protective factors against this bias and discuss the impact of these effects on clinical and research applications of auditory-perceptual voice examinations.

BACKGROUND

Voice Disorders

A voice disorder is defined by the American Speech-Language-Hearing Association (ASHA) as “when [vocal] quality, pitch, and loudness differ or are inappropriate for an individual's age, gender, cultural background, or geographic location,” (American Speech-Language-Hearing Association, n.d). This encompasses a wide variety of symptoms, etiologies, and functional impacts. The prevalence of voice disorders among individuals in the United States is between 0.26–3.2% (Best & Fakhry, 2010). In addition, there is a much higher prevalence in certain occupations with high vocal use such as salespeople, teachers, singers or actors, and manufacturing or factory workers. For example, the lifetime prevalence of voice disorders in teachers is 57.7% compared to a lifetime prevalence in the general population of 28.8% (Roy et al., 2004), although it should be noted that the definition of voice disorders used in Roy et al. (2004) was broader than the definition used in other studies. Voice disorders are also more common among adult females than males (Roy et al., 2005). The prevalence and impact of voice disorders make it an important subset of speech and language pathology’s scope of practice.

Voice disorders can also have a dramatic impact on the physical, emotional, and mental wellbeing of an individual. Merrill et al. (2011) found a significant association of voice disorders with incidents of anxiety and depression as well as associations with overall physical and emotional role limitations. Those with voice disorders also reported lower social functioning and overall health than the general population (Merrill et al., 2011). Continued research in this area is needed to learn how best to serve these individuals.

The components of a comprehensive voice evaluation according to Reghunathan & Bryson (2019) include laryngoscopy, videostroboscopy, patient questionnaires, an auditory-perceptual evaluation, and acoustic and aerodynamic measures. This paper will be focused on the auditory-perceptual evaluation. Common scales used clinically and in research include the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) (Kempster et al., 2009) and Grade Roughness Breathiness Asthenia Strain scale (GRBAS) (Hirano, 1981).

Auditory-perceptual Examinations

There is a need for continued research into the reliability of auditory-perceptual measures and the factors that can affect listener reliability and judgements. Because auditory-perceptual evaluations are inherently a subjective measure, inter-rater agreement is low (Zraick et al., 2011; Kelchner et al., 2010; Karnell et al., 2007), yet it is important for progress tracking, management, and inter-clinic communication to have a reliable measure of voice perceptual qualities. There is a wealth of research into various aspects of perceptual judgements, including research on the type of scale used (Kreiman et al., 2007b), which aspects of the CAPE-V have the best reliability (Zraick et al., 2011; Kelchner et al., 2010; Karnell et al., 2007), listener experience and its role in reliability (Eadie et al., 2010), training of listeners (Eadie & Baylor, 2006), and prior knowledge of diagnosis (Eadie et al., 2011). Many of these studies concluded that raters are swayed by various factors. For example, Eadie et al. (2011) concluded that a rater's knowledge of a diagnosis influenced their rating of the voice, and Zraick et al. (2011) found that CAPE-V percepts differ in their judgement reliability. However, to the best knowledge of this author, there is a lack of research on perceptual bias in ratings of voice quality.

Despite these weaknesses enumerated above, auditory-perceptual evaluations are an important part of comprehensive voice evaluations. Some strengths are posited in Oates (2009), including that voice is experienced as a perceptual response to an acoustic stimulus. Because of this, there is a potentially inherent understanding of perceptual qualities of voice for clients and clinicians alike (Kreiman et al., 1993). For example, describing a voice as “breathy” is more easily understood by patients and physicians than a report of the noise-to-harmonics ratio. There is also an ease in administration for both patient and rater as they require no specialized equipment and can be completed quickly (Oates, 2009). They are a noninvasive measure and can be repeated frequently with no fear of adverse effects (Webb et al., 2004). Additionally, there is a lack of consensus on the correlation of objective acoustic measures with perceptual dimensions. Several studies have demonstrated only low to moderate predictive relationships for acoustic and perceptual measures (Ma & Yiu, 2006; Kempster et al., 1991; Yiu, 1999). For these reasons, instrumented measures cannot replace a perceptual evaluation, so it is imperative that every factor contributing to potential listener/rater bias be examined.

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

The CAPE-V was developed in 2009 as part of ASHA’s Special Interest Division 3, Voice and Voice Disorders, as an attempt to produce a standardized method of completing voice auditory-perceptual evaluations. It was sponsored by the aforementioned ASHA Special Interest Group 3 and the University of Pittsburgh. The CAPE-V incorporates six voice percepts (overall severity, roughness, breathiness, strain, pitch, and loudness) on a 100-mm visual analog scale (VAS). The VAS incorporates non-equally spaced anchors of mild, moderate, and severe for each voice perceptual quality, as well as additional lines for other voice qualities judged to be important to record on a patient-by-patient basis such as comments on resonance, diplophonia, or glottal fry. CAPE-V is commonly used as an outcome measure for treatment, as well as a diagnostic tool (Behrman, 2005; De Bodt et al., 1996).

Sequential Effects

Sequential effects are a type of perception bias. It is the concept that our perception of a stimulus is influenced by the context in which it is experienced. This occurs in judgements of attractiveness (Pegors et al., 2015), pitch (Arzounian et al., 2017), loudness (Kind & Lockhead, 1981), and even essay grading (Attali, 2011; Zhao et al., 2017). There are two main types of sequential effect, contrast bias and assimilation bias.

Contrast bias occurs when one stimulus's rating is conversely affected by the previous stimulus, i.e., the rating moves in the opposite direction of the previous stimulus. For example, Kenrick & Gutierres (1980) showed photographs of women to men who were watching *Charlie's Angels* and men who were watching another show. The men watching *Charlie's Angels*, a show with three traditionally beautiful women as the main characters, rated the photographs as less attractive than those watching another show. This demonstrates a contrast bias because the comparison enhances the differences between the two.

Assimilation bias is the opposite effect; the stimulus preceding has a convergent effect on the rating of the next stimulus. In other words, the second stimulus will be rated more closely to the preceding stimulus. For example, in Schwarz & Bless (1992) when participants were primed with Richard Nixon, the overall trustworthiness ratings of politicians as a whole decreased. Comparison to Nixon, an accepted standard of an untrustworthy politician, had a negative effect on sequential ratings of politician trustworthiness, like an after-glow effect. These two types of sequential judgement effects can and do co-occur as in Pegors et al. (2015) in which both a contrast bias and assimilation bias were shown on judgements of attractiveness. Often one has a stronger effect than the other, as shown in the Pegors et al. (2015) study.

There are various theories as to why one type of bias might have a stronger effect than the other. One theory put forth in Mussweiler (2003), termed “The Selective Accessibility Mechanism”, hypothesizes that the initial selection of a comparison item made by a judge will dictate the effect outcome. If the comparison and the target are fundamentally similar, an assimilation bias will occur because the judge is searching for similarities to complete their judgement. The opposite effect would then occur if the target and comparison are fundamentally different, i.e. a contrast bias would result. These initial comparisons depend on the accessibility of a standard and could theoretically be manipulated to produce different types of biases.

In theory, the contextual presentation of voices should also be susceptible to these effects. By presenting moderately dysphonic voices in the context of severe or mildly dysphonic voices, the most easily accessible standard would be the voice immediately preceding. This can also be extrapolated to predict that in auditory-perceptual voice evaluations by raters with experience rating voices, the rater should have a stronger internal standard (i.e., have the experience to know a moderately dysphonic voice) and therefore be less influenced by the easy availability of a skewed standard than novice raters.

Assimilation bias is considered the default judgement mechanism because judgements resulting in assimilation bias are made more quickly, and judges are more likely to first consider similarities between the target and comparison (Chapman & Johnson, 1999; Lockwood & Kunda, 1997; Miyake & Zuckerman, 1993; Nosanchuk & Erikson, 1985). This is also supported by Martin et al. (1990) in which they conclude that assimilation bias is more likely when the raters lack motivation or when cognitive resources are taxed, both situations in which the easier or default judgement would be likely to occur.

The Selective Accessibility Mechanism theory also explains why anchors can influence the effects of sequential bias as shown in Sherif & Hovland (1961). When the given anchor was close to the test stimulus, assimilation bias occurred. When the anchor was far from the given stimulus, contrast bias occurred. This is because a given anchor is necessarily the most accessible standard for the judge because they are directed to use the anchor as their standard, so each stimulus will be compared to it. This decreases the likelihood that a stimulus is judged compared to the one preceding it. Experienced raters of auditory-perceptual evaluations should have an internal anchor to counteract these sequential effects, but the strength of it is likely influenced by individual factors such as a willingness to seek the internal anchor instead of relying on easily accessible context, as Martin et al. (1990) suggests.

Different raters are also affected in various ways by these types of biases, although the literature does not agree on the direction of this influence or what key factors may influence a rater. For example, Pelham & Wachsmuth (1995) conclude that assimilation bias is more likely when a rater is confident in their choices, i.e., when a rater is more experienced. In contrast, Attali (2011) found that rater quality (defined as having a high correlation with other raters' scores) was a protective factor against assimilation bias; and although he found an overall assimilation effect in sequential judgements of essay scoring, 4% of the raters showed no sequential effects at all. The paper did not explore this further, but it suggested that there are individual differences in susceptibility to sequential effect biases. What those differences are remains somewhat of a mystery. Some potential factors that will be explored in this experiment include amount of experience, time taken to complete the experiment, and confidence level in ratings of voice quality.

This thesis aims to examine sequential effects in ratings of voice quality. Two groups of raters will be asked to judge only the overall severity rating using the CAPE-V for a series of voices containing the moderately dysphonic test stimuli in three contexts: neutral, severe, and mild. One group is composed of speech language pathologists (SLP) with experience rating voices, the other of SLPs who work with adults but not in the voice domain. A brief survey at the end of the experiment will aim to capture potential individual protective factors against sequential effects.

Specifically, this project intends to answer the following questions:

1. Are there sequential effects on the overall severity rating of the CAPE-V by raters with and without experience rating voices?

2. If so, is there a contrast or assimilation bias, and do the effects differ by rater experience in size or direction?

3. Finally, what factors protect or mediate a rater's susceptibility to sequential effects?

First, the expectation is that sequential bias effects will occur for raters with and without experience rating voices. It is hypothesized that raters with experience rating voices will be less susceptible to sequential effects because of their internal anchors and therefore have smaller effects, regardless of direction. They are also predicted to show an assimilation bias because they are likely to be confident in their judgements and to complete the task more quickly, which would lead to a default judgement mechanism or assimilation bias. Raters without experience rating voices, however, do not have a strong internal standard and are therefore expected to show a contrast bias because they will be susceptible to sequential effects due to the previous stimuli. Finally, some factors that are expected to protect against sequential effects are greater experience rating voices, more confidence in rating voices, and a longer amount of time taken to complete the task.

METHODS

Participants

Raters with and without experience rating voices were participants in this study. The group of raters with experience rating voices was composed of certified SLPs with a minimum of three years of experience in the field including a year of Clinical Fellowship, e.g., an SLP in their third year of practicing, but only their second year with their CCCs, would qualify as an expert rater. They were required to have previous experience scoring the CAPE-V and voice clients as a minimum of 50% of their caseload in the previous year. The group of raters who were inexperienced in rating voices was composed of SLPs who do not work in the voice field, but engage in clinical practice with an adult population. They were also required to have a minimum of three years of experience. They had less than a year of working with voice clients, fewer than 5% of their caseload as voice clients across the last year, and little or no experience using the CAPE-V clinically. Each group contained 10 participants for a total of 20 raters. An effort was made during recruitment to match the two groups by overall experience as an SLP. The average years in the field for participants without experience rating voices was 12.8 years, compared to an average of 14.9 years for the experienced in voice ratings group. Due to the remote nature of the experiment, the participants hailed from a wide geographic area including but not limited to Washington, California, Utah, Michigan, Massachusetts, and Rhode Island. The participants were compensated for their time.

Stimuli

Each voice stimulus was composed of one instance of /a/ and one instance of /i/ for two to four seconds each, and sentences four through six of the Rainbow Passage (Fairbanks, 1960). The stimuli were selected from a pre-existing database to encompass a wide range of severity levels. Then they were sorted into mild, moderate, and severe groups based on overall severity ratings of dysphonia collected using the CAPE-V. These ratings were established by averaging two ratings made by SLPs with significant experience rating voices on the CAPE-V. Inter and intra-rater reliability were calculated using Pearson's r. Interrater reliability was 0.89, interpreted as a strong relationship and therefore good reliability (Akoglu, 2018). Intra-rater reliability for Rater 1 was 0.90 and 0.94 for Rater 2, again suggesting a strong relationship and good reliability (Akoglu, 2018). These categories are displayed in Table 1.

Voices between each category were eliminated in order to more fully distinguish in and out of group voices. Particularly unique voices as determined by the author, such as those with a strong accent, errors in reading, or co-occurring resonance or articulation issues, were excluded. This is because particularly unique voices are more likely to be remembered and participants may then rate a voice the same based on memory rather than their current perception. The diagnoses represented included muscle tension dysphonia, vocal fold nodules, primary vocal hyperfunction, adduction laryngeal dystonia, granuloma, edema, and vocal fold paralysis. These are all diagnoses that are common for voice pathologists to assess, and individuals with these disorders do not tend to have co-occurring resonance or speech deficits. To further reduce variability across the stimuli, all voices selected were female and adult. The average age was 38.8 (SD= 16.6) ranging from 19–76. Four of the moderate voices were selected to be the test stimuli set. These voices were taken from the middle of the moderate test set based on the previous CAPE-V ratings and are also summarized in Table 1.

Table 1: Distribution of stimuli by overall severity of dysphonia CAPE-V rating

Group	**CAPE-V Range**	**Mean**	**Standard Deviation**
Mild	5–7	6.1	0.6
Severe	53–89	68.1	13.4
Test Set	22–25	23.1	1.4

Note: CAPE-V = Consensus Auditory Perceptual Evaluation of Voice

The stimuli were edited using PRAAT software (Boersma & Weenink, 2021) so the order of each was the same: one second of silence, a sustained /a/, one second of silence, a sustained /i/, one second of silence, and sentences 4–6 of the Rainbow Passage. The stimuli were normalized for peak amplitude using the “normalize” function in the PRAAT Voice software package. This is intended to reduce variability within the voices and ensure that each stimulus is played at the same volume.

Design

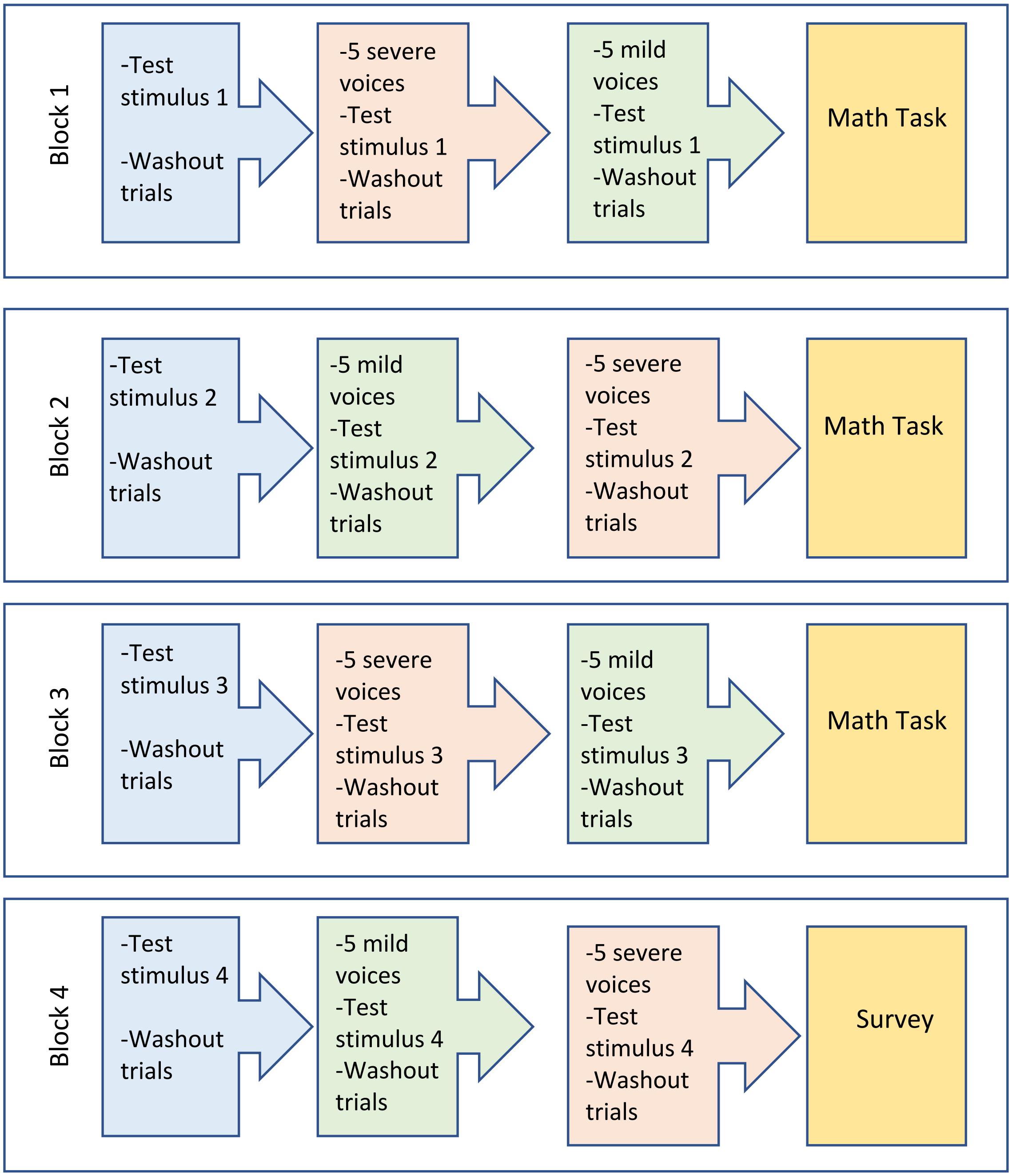
The independent variable was the context in which the stimuli were presented: neutral, severe, or mild. These contexts were composed of the previous stimuli being none, five severely dysphonic voices, or five mildly dysphonic voices, respectively. The dependent variable was the overall severity of dysphonia judgement made by the rater of the moderate test stimuli voices using the CAPE-V.

The experiment was composed of four blocks separated by a math task. All three conditions were present within each block: neutral, severe, and mild. The neutral condition contained the test stimulus only. The severe and mild conditions contained five severe or mild voices then the test stimulus. Between each condition, washout trials of four voices of variable severity were presented. The experiment flow is visualized in Figure 1. This layout ensured that the test stimuli, except in the first neutral instance, was heard after nine other voices, the last five of which create either a severe or mild environment. Each block consisted of 25 voices total.

Between each block, a mandatory approximately two-minute break was implemented which consisted of a simple math task of whole number addition, subtraction, multiplication, and division. Participants were not allowed to use a calculator for this task. The purpose of the mental math task was to prevent the participants from reflecting on their ratings and therefore remembering the voices and their ratings. In addition, it served to extend the time between blocks in an effort to establish a neutral condition once more for the test stimulus of the following block. This strategy was based on the findings of Attali (2011) which found that sequential effects on essay ratings decreased in strength with the amount of time between ratings.

The full experiment consisted of four of these blocks, each with a different test stimulus for a total of 100 ratings. In each block, the washout trial voices were randomized in order. The five severe and mild condition voices were also randomized in order for each block. Within each block, every participant heard the neutral condition first. Half the participants of each group then heard the mild condition followed by the severe condition, whereas the other half had the opposite order of listening.

Figure 1: Diagram of a sample experiment flow depicting order of stimuli



Procedure

This experiment was conducted remotely on the web-based platform Gorilla (www.gorilla.sc). All participants opened the Gorilla experiment using Google Chrome browser from an individual participant link. Before beginning the experiment, each participant met via Zoom (Zoom Video Communications, Inc., 2020) with the author to complete informed consent and receive directions. The participant and author then turned off their video and sound to complete the ratings portion of the experiment but stayed on Zoom in case questions arose or any unexpected incident occurred. One incident was reported in which a participant's browser crashed while loading the experiment. No ratings had yet been made and the participant was instructed to restart the browser, at which point they were able to successfully load and complete the task.

Participants were instructed to use headphones to reduce variability among the raters, and a screening to ensure headphone use was performed prior to beginning the rating task. This entailed a volume setting task and a headphone screening. In the volume setting task, participants heard white noise and adjusted the white noise volume to a comfortable level. This volume was then used for the entirety of the experiment. The goal of the volume setting task and normalization for volume of the stimuli is to reduce variation within the experimental paradigm. The headphone calibration task was drawn from the Gorilla open-source material repository to ensure that raters used headphones for the experiment (Milne et al., 2020). It consisted of three segments of white noise, one with a faint tone underneath the white noise. Raters were asked to identify which segment contained the tone. This headphone screening is 80% effective at detecting those who are not wearing headphones (Milne et al., 2020). Five participants failed the first headphone screening and were instructed to try again with a different set of headphones. All passed the on second try and successfully completed the experiment.

Ratings of the overall severity of dysphonia were collected for each voice. Overall severity was chosen because research shows it has the highest interrater reliability of all aspects of the CAPE-V (Solomon et al., 2011, Karnell et al., 2007, Kelchner et al., 2010) and is therefore the most stable. This is important because a more stable rating is less likely to show effects due to chance, so the effects seen on this overall severity rating are likely due to the manipulated condition and not complicating factors. The rating scale consisted of a bar with MI (mild) ME (moderate) and SE (severe) markers at non-equally spaced locations to prevent end effects in accordance with the CAPE-V rating document available from ASHA (ASHA CAPE-V Form. (n.d.)). Participants were instructed to click on the bar and then drag the marker to the spot they judged to be the most accurate rating. The program then recorded a rating as a whole number from 0–100 as measured from the left end of the scale. Participants were not able to see the whole number correspondence during the task. The participant view is visualized in Appendix A. This was designed to mimic the official CAPE-V administration guidelines from ASHA that direct the rater to make a tick mark on the line and then measure from the left end in millimeters to determine the numerical rating.

Each participant made 100 total ratings, consisting of four blocks of 25 voices, each containing the three conditions. In between blocks, a two-minute math task was given. The average time for completion was 53.7 minutes ranging from 36 to 71 minutes.

After rating all the stimuli, participants completed a brief survey. This survey was intended to capture some potential protective factors against sequential effects. The survey included a 100-point VAS of how confident participants were in their ratings, the length of time they had been a practicing SLP, and the length of time they had been using the CAPE-V in their clinical practice. A full list of survey questions can be found in Appendix B. The answers to these questions, along with the amount of time taken to complete the experiment, were compared to the sequential effect of each rater to determine which, if any, factors influenced the size of the sequential effect.

Following the post-experiment survey, participants returned to the Zoom meeting and were debriefed by the author. The debrief included a short explanation of the experiment’s purpose and objectives. Appendix C contains diagrams of the full experimental flow process from Gorilla.

Statistical Analysis

To test the hypothesis that sequential effects would occur for SLPs with and without experience rating voices, a mixed effects analysis of variance (ANOVA) was conducted with fixed factors of group assignment (experienced/inexperienced in voice ratings), test stimulus, and condition (neutral, mild, severe), and with rater as a random factor. The ANOVA also included interactions of group and condition and test stimulus and condition. Effect sizes were calculated with partial eta squared (η_p^2) and interpreted according to guidelines suggested by Mile & Shevlin (2001).

A Dunnett's *post hoc* test was conducted, given significant findings from the ANOVA, to test the hypotheses that SLPs with experience rating voices will demonstrate an assimilation bias, and SLPs without experience rating voices will demonstrate a contrast bias. The control for the Dunnett's test was the neutral condition, which was compared with the mild and severe conditions. Effect sizes were calculated using Cohen's *d* and interpreted according to guidelines in Cohen (1988).

To address the final hypothesis that the following factors—years as an SLP, years using CAPE-V, confidence in general using the CAPE-V, confidence in ratings today, time taken to complete the ratings, prior voice training, and prior experience rating voices—influence the size of bias demonstrated by each individual rater, a multiple linear regression was conducted. This assessed the aforementioned individual factors and their potential impact on the difference between ratings in neutral and severe conditions for each rater, since the Dunnett's *post hoc* showed a significant effect of the severe condition only. First, the average difference of severe condition ratings minus the neutral condition rating for the four test stimuli was calculated for each rater. The linear regression then incorporated the continuous variables of years as an SLP (including CFY), years using the CAPE-V, confidence in general using the CAPE-V on a 100-point VAS, confidence today in ratings on a 100-point VAS, and time in minutes to complete the study, as well as the categorical variables of group assignment and participation in any voice training after high school. All statistical analyses were conducted in Minitab (Version 19, Minitab Inc., State College, PA).

Finally, interrater reliability was calculated using an intraclass correlation coefficient (ICC). The stimuli used in this calculation were part of the washout trials between each condition which included mild, moderate, and severely dysphonic voices. Each participant rated these 12 stimuli a total of four times. The first rating of each stimulus was used in the interrater reliability calculation. The ICC was conducted based on a mean-rating (k=20), absolute agreement, two-way mixed effects model.

All four ratings of the same 12 stimuli were used to calculate intra-rater reliability for each rater, again using ICC, based on an absolute agreement, two-way mixed effects model. These reliability analyses were completed in Microsoft Excel (Version 16.55, Microsoft Corporation, 2021). Reliability was interpreted in terms of poor, moderate, good, and excellent reliability using guidelines suggested by Koo & Li (2016).

RESULTS

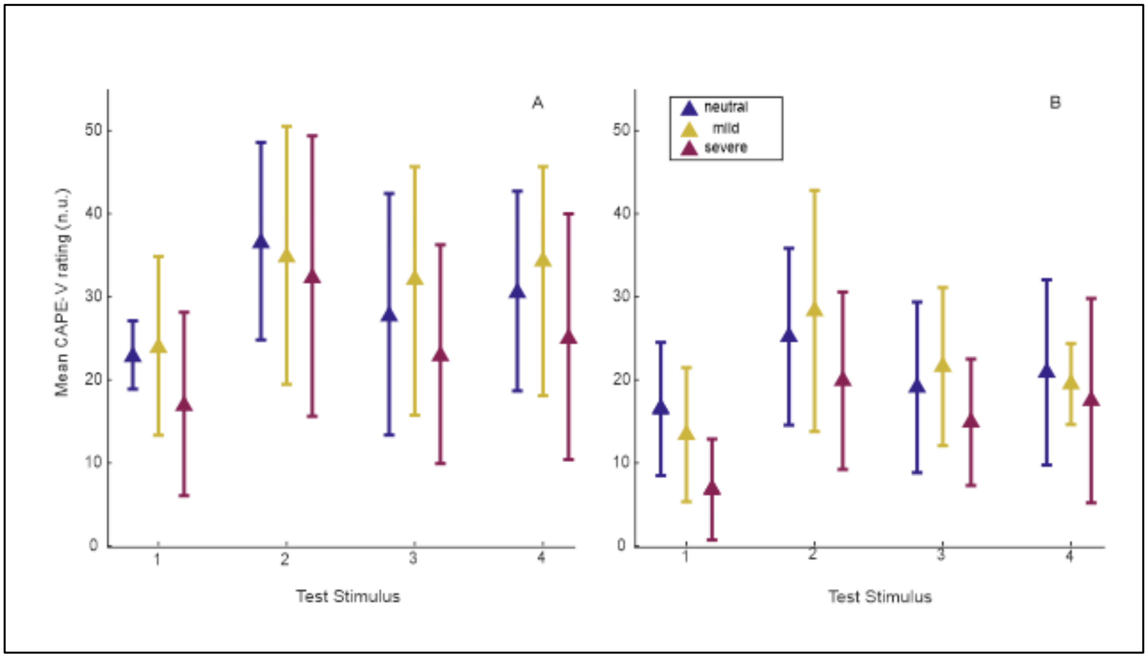
Participants rated voices on overall severity of dysphonia using the CAPE-V 100-point VAS in three different conditions: neutral, mild, and severe. The ratings from each condition for four test stimuli were compared to assess sequential effects.

Sequential Effects

The test stimuli were chosen because they were rated by two expert raters in the center of the moderate array of voices with average ratings of 22.1, 22.4, 22.4, and 25.5 using the CAPE-V. The study participants rated the test stimuli voices similarly with average ratings in the neutral condition of 19.7, 23.5, 25.8, and 31.0, respectively.

In general, SLPs who were experienced in rating voices rated all samples, regardless of condition, on average 10 points more severe than inexperienced raters. The experienced raters group also had more variation in ratings than the inexperienced raters group, as evidenced by larger 95% confidence intervals depicted in Figure 2 below. The means and 95% confidence intervals of CAPE-V ratings for each test stimulus are illustrated for each group by condition in Figure 2.

Figure 1: Means and 95% confidence intervals for experienced in voice ratings group (Panel A) and inexperienced in voice ratings group (Panel B) for CAPE-V ratings in neutral, mild, and severe conditions.



In the mixed effects ANOVA, there was a statistically significant effect of rater group and of condition, but not of interactions between group and condition or test stimulus and condition. There was a medium-large effect of group assignment on CAPE-V ratings ($F(1, 207) = 8.19$, $p<0.05$, $\eta_p^2 = 0.23$). There was also a medium effect size of condition on CAPE-V ratings ($F(2, 207)=10.41$, $p<0.05$, $\eta_p^2 = 0.09$). Full results can be found in Table 2.

Table 2: Results of mixed effects analysis of variance for effect on CAPE-V ratings with fixed factors of group assignment, speaker, and condition, and with rater as a random factor.

Factor	***df***	***F***	***p***	***n_p^2***	***Effect size***
Group assignment	1	8.2	.01	0.2	Large
Test stimulus	3	18.1	< .001	-	-
Condition	2	10.4	< .001	0.1	Medium
Rater	18	7.8	< .001	-	-
Group x Condition	2	0.2	-	-	-
Test stimulus x Condition	6	0.3	-	-	-

Note: Significant at p<0.05 (nonsignificant values dashed out); effect sizes: 0.1 = small; .06 = medium; .14 = large (Miles & Shevlin 2001).

Dunnett's *post hoc* testing between the neutral condition and severe/mild conditions found significantly lower ratings, on average 6 points, of the overall severity of dysphonia for the severe condition only (p_{adj}<.001, *d* = 0.75). This is a large statistically significant contrast bias for all raters in the severe condition only. Full results are reported in Table 3.

Table 3: Results of Dunnett's post hoc test for difference in CAPE-V ratings between conditions with neutral condition as control group.

Condition	***p***	***d***
Mild	0.7	-
Severe	0.0	0.8

Note: Level of significance; p<0.05; Cohen's *d* effect sizes; 0.2 = small, 0.5 = medium, 0.8 = large (Cohen, 1988).

Individual Factors

The post experiment survey collected information for each participant including time as an SLP, time using the CAPE-V, and various confidence ratings. Of the experienced at voice ratings group, seven participants reported voice training post-high school ranging from 1–20 years. In the inexperienced at voice ratings group, only one participant reported voice training of four years in length. In general, SLPs with experience rating voices had higher confidence, both in general using the CAPE-V and following the rating task, as well as required less time to complete the task. Full descriptive statistics of these individual factors are reported in Table 4.

Table 4: Descriptive statistics of raters organized by group.

Individual Factor	**Experienced at voice ratings**		**Inexperienced at voice ratings**	
	Mean	**Range**	**Mean**	**Range**
Years as SLP	14.9	3–37	12.8	3–39
Years using CAPE-V	10.4	3–18	0.2	0–1
Confidence in general	87.2	55–100	21.2	1–44
Confidence today	80.3	50–100	32.4	6–94
Time (mins)	50.9	36–71	56.5	37–87

Note: SLP = speech language pathologist; CAPE-V = consensus auditory-perceptual evaluation of voice.

The overall multiple linear regression was statistically significant (R^2 = 0.67, F(7, 12) *F*= 3.51, *p* < .05). Confidence in general using the CAPE-V significantly predicted the difference in neutral and severe rating conditions (β =

0.470, p = 0.021), as well as confidence in ratings today (β = -0.380, $p < .05$) and group assignment as a proxy for experience rating voices (β = 18.00, $p < .05$). Confidence in general and little or no experience rating voices had the effect of increasing contrast bias because the β value is in the positive direction. Confidence in ratings today caused a reduction in contrast bias because the β value is negative.

Years as an SLP, years using the CAPE-V, time to complete ratings, and prior voice training did not significantly predict a difference in ratings between neutral and severe conditions. Full results can be found in Table 5.

Table 5: Results of multiple linear regression for difference between neutral and severe condition overall severity of dysphonia CAPE-V ratings with years as SLP, years using CAPE-V, confidence in general, confidence in ratings today, and time as continuous predictors and with group assignment and prior voice training as categorical predictors.

Individual Factor	**β**	***p***
Years as SLP	0.3	0.1
Years using CAPE-V	0.3	0.5
Confidence in general	**0.5**	**0.0**
Confidence today	**-0.4**	**0.0**
Time (mins)	0.1	0.4
Prior voice training	4.9	0.2
Group assignment	**18.0**	**0.0**

Note: Level of significance; p<.05 (significant differences bolded in table), β =standard beta coefficient; SLP = speech language pathologist; CAPE-V = consensus auditory-perceptual evaluation of voice.

Interrater reliability

Interrater reliability is a measure of the agreement raters had with one another for the same set of stimuli. The overall interrater reliability computed using ICC for all 20 participants was 0.62, interpreted as moderate reliability (Koo & Li, 2016). The interrater reliability for the experienced at rating voices group (n=10) was 0.59, indicating moderate reliability, whereas the interrater reliability for the inexperienced at rating voices group (n=10) was 0.71, again interpreted as moderate reliability (Koo & Li, 2016).

Intra-rater reliability

Intra-rater reliability is a measure of how consistent an individual rater is with themself across multiple ratings for the same stimulus. The average intra-rater reliability for all 20 participants was 0.83, which indicates good reliability (Koo & Li, 2016). The experienced at voice ratings group had an average intra-rater reliability of 0.85, interpreted as good reliability (Koo & Li, 2016). The average intra-rater reliability of the inexperienced at rating voices group was 0.80, again indicating good reliability (Koo & Li, 2016).

DISCUSSION

Effect of condition

A contrast bias was found for both groups of raters in the severe condition only. This means the moderate test stimuli were rated on average as six points less severe when heard after severe voices than when rated on their own in a neutral context. This is a contrast bias because the voice ratings decreased after exposure to more severe voices. This is not a clinically significant difference according to Eadie & Kapsner-Smith (2011), where a difference of more than 7.14mm or approximately seven points on the CAPE-V is considered a clinically significant difference. Contrary to the proposed hypothesis, the contrast bias was found for both groups of raters. It was expected that the experienced at rating voices group would demonstrate an assimilation bias instead.

There was not a statistically significant difference, and therefore no sequential effect, in ratings of overall dysphonia between the neutral and mild conditions. One possible explanation is the distribution of test stimuli used. Although the test stimuli were drawn from the middle of the data set of dysphonic voices, they were not in the middle of the CAPE-V 100-pt scale, i.e., with ratings of ~50, because the data set of dysphonic voices was clustered to the mild end of CAPE-V ratings. This resulted in a smaller difference on the CAPE-V rating scale between the mild condition voices (mean = 6.6) and the test stimuli (mean = 21.6) than the difference in the severe condition voices (mean = 55.7) and the test stimuli.

This lack of sequential effect in the mild condition could also be explained through the Selective Accessibility Mechanism framework (Mussweiler, 2003). This framework proposes that the perceived difference between the target and comparison dictates the type of sequential effect that will result. In this study, the target or “test stimulus” was close to the comparison or “mild condition”, and there was no sequential effect observed, perhaps because the distance between the two was not large enough to find either a similarity or difference between the two that would then result in a sequential effect.

Comparison to Previous Studies: Size of Sequential Effect

It can be reasoned that the size of the sequential effect found will depend on the rating scale used. For example, a difference of 1-point on a 10-point scale is a much larger difference, a 10% difference, than a difference of 1-point on a 100-point scale which would be a change of 1%. To compare the sequential effect sizes in different studies, the percent change in the scale the sequential effect had was calculated by dividing the effect size found by the scale size and then multiplying by 100. The actual effect size numbers are reported in Table 6.

The percent change in this study was 6%, meaning that the average sequential effect found was 6% of the scale used. This was within the range of percent change found in Schwarz & Bless (1982) which ranged from 2.5–15%, and the range of percent change found in Zhao et al. (2017) of 2.9–8.9%. Daly & Dickson-Markman (1982) found a larger percent change and therefore a larger and more meaningful sequential effect size with a range of 8.5–18.4%. Although the contrast bias was not clinically significant (Eadie & Kapsner-Smith, 2011), overall, the sequential effect size in this study was comparable to previous findings, despite the difference in scale sizes and tasks.

Table 6: Size of sequential effects reported in the literature.

Publication	**Sequential effect size**	**Scale**	**Percent change**
Daly & Dickson-Markman 1982	0.6–1.3	7-point	8.5%–18.4%
Schwarz & Bless 1982	0.3–1.8	12-point	2.5%–15%
Zhao et al 2017	0.6–1.8	20-point	2.9%–8.9%

Note: Percent change calculated as (sequential effect size/scale).

Effect of experience

Participants with experience assessing voices rated voices as more severe, by an average of 10 points, than those who were inexperienced at rating voices, regardless of the condition. This suggests that the difference in group membership, i.e., experience rating voices, has both a statistically and clinically significant effect on voice ratings of overall severity of dysphonia using the CAPE-V (Eadie & Kapsner-Smith, 2011). This indicates that working with voice clients as a majority of one's caseload (>50%) results in more severe voice ratings, but not a different sequential effect direction because there was no interaction between group and condition shown in the mixed measures ANOVA. It was expected those with experience rating voices would have an assimilation bias, but a contrast bias was found for both groups instead. Although both groups showed a contrast bias, the bias was smaller in experienced voice raters.

Since group assignment as a proxy for experience rating voices had the largest coefficient in the linear regression, experience rating voices had the largest effect on the size of the contrast bias observed in the data. Those in the inexperienced at voice rating group had an 18-point increase in the average difference of test stimuli ratings in the neutral vs severe conditions, meaning the contrast bias observed is larger for inexperienced raters. This supports Mussweiler (2003) who posits that increased experience results in a stronger internal anchor and therefore less susceptibility to sequential effects. So although both groups demonstrated a contrast bias by rating moderate voices as more severe when they followed severe voices, the rating change by experienced voice clinicians was smaller than the change made by those who were inexperienced at rating voices.

In contrast, it was found that years as an SLP, years using the CAPE-V, and prior voice training did not significantly impact the size of the contrast bias observed. It is interesting that time using the CAPE-V did not have a significant effect, whereas group assignment did. This suggests that if the experience level of a clinician is greater than 50% of a caseload of voice clients across 3 years, additional time using the CAPE-V does not result in a smaller contrast bias. Total experience as an SLP did not protect against sequential effects in this study either—suggesting that only experience in the voice domain has a significant effect on susceptibility to these sequential effects.

The role of listener experience in voice ratings continues to be a debated subject in the field. The findings of this study are in opposition to Helou et al. (2010), which found that inexperienced listeners rated voices as more severe than experienced listeners. In contrast to both this study and Helou et al. (2010), Eadie & Kapsner (2011) and Eadie et al. (2010) found no significant effect of experience on ratings of voice quality. These conflicting results could in part be explained by the varied definitions of “experienced” and “inexperienced” raters used in each study. For example, this study defined experienced raters as those with 3+ years as an SLP with >50% of a voice caseload, but Helou et al. (2010) recruited SLPs and ENTs with no defined minimum voice experience level and contrasted their judgements with non-voice medical professionals. Both Eadie & Kapsner (2011) and Eadie et al. (2010) categorized an experienced listener as one with 6+ years of voice experience and their inexperienced group was laypeople without prior voice knowledge or experience. Due to this, it is difficult to draw any overarching conclusions about the effect of experience on voice ratings.

Effect of Rater Confidence

Both confidence in general using the CAPE-V and confidence in ratings made “today” had a statistically significant effect on the size of contrast bias observed. The coefficient of confidence in general was 0.47, meaning a 1-point increase in confidence rating voices in general translates to a 0.47-point increase in the rating difference, or a larger contrast bias. Greater confidence in general using the CAPE-V assessment tool could correlate with less motivation during the task because the rater is less likely to rethink or double check their initial rating. This aligns with Martin et al. (1990) which suggests that less motivation will result in a larger sequential effect because they are less likely to seek the internal anchor and instead rely on easily accessible context, i.e., the voices preceding.

In opposition, the contrast bias was observed to be smaller with increased confidence in ratings “today”, i.e., directly after the ratings were collected, with coefficient of -0.38. This indicates that a 1-point increase in confidence today translates to a 0.38 decrease in rating difference, or a smaller contrast bias. Greater confidence today then is a protective factor which reduces the size of sequential effect observed. This likely reflects increased motivation during the task as participants with higher confidence in ratings “today” feel they were successful in carefully considering each voice and rating them appropriately. The participants’ answers are reflecting their confidence in sequential voice ratings as opposed to confidence using the CAPE-V tool itself, which could lead to decreased motivation in this particular task as discussed above.

Although these effect sizes seem small, the cumulative nature of the effect could have a large impact on the size of contrast bias observed in individuals. For example, a participant with a confidence in general rating of 98 would have approximately a 41-point larger contrast bias than a participant with a confidence in general rating of 10. These results indicate that rater confidence can be a key factor in voice rating reliabilities, and more research into this particular factor is indicated as the full extent and nature of this interaction was not within the scope of this project.

Comparison to Previous Studies: Reliability

The intra-rater reliability findings in this study were aligned with findings from previous studies on CAPE-V reliability for the overall severity rating. Kelchner et al. (2010) and Karnell et al. (2007) found good reliability. Nemr et al. (2012) found excellent reliability interpreted according to Koo & Li (2016). Zraik et al. (2011) found moderate reliability (Akoglu, 2018). The findings of this study are consistent with these studies, as it also found good reliability according to Koo & Li (2016) with an ICC of .83 for all raters.

Interrater reliability has historically been lower than intra-rater reliability in studies on the CAPE-V's reliability, and this study's findings followed this trend. The ICC interrater reliability of this study was .623, interpreted as moderate reliability according to Koo & Li (2016). In Kelchner et al. (2010), they found highly variable interrater reliability ranging from moderate to good. Nemr et al. (2012) found excellent reliability and was the best interrater reliability found among these studies. Zraik et al. (2011) and Karnell et al. (2007) both found good reliability (Akoglu, 2018).

Overall, this study's interrater reliability findings were on the lower end of the previously reported findings. A possible explanation for this could be the wide geographic range of the participants. SLPs within a single clinic are likely to have more cohesive ratings and higher reliability because they have a collective knowledge of what they consider to be mild, moderate, and severely dysphonic voices. They are also likely to have undergone the same training. Of the studies mentioned above, all except Zraik et al. (2011) used raters from one facility. Zraik et al. (2011) reported participating raters from 17 different facilities; similarly, this study had participants from 14 different facilities. The interrater reliability of Zraik et al. (2011) was most closely aligned with the results of this study, and on the lower end of ranges found in the literature, supporting the conclusion that a wide geographic spread of raters can result in lower interrater agreement.

Previous literature on the interaction of CAPE-V reliability and experience levels of voice raters is inconclusive, and this study follows that trend. Kreiman & Gerratt (2000) found that those with experience rating voices have more variability, and therefore less interrater reliability, in their ratings than those without. They suggest this is due to experienced listeners being more finely tuned to the various perceptual qualities of voices and using a flexible strategy to incorporate this knowledge into judgements, thereby decreasing their reliability. That is a possible explanation for this study's findings of interrater reliability of those with experience rating voices being lower than those without—0.59 and .71, respectively. This is in opposition to Helou et al. (2010) which found greater interrater reliability among raters with experience rating voices compared to those without. In contrast, Eadie et al. (2010) found no significant difference in interrater reliability between experienced and inexperienced raters.

These findings indicate that there is no single experience level at which it can be predicted that raters will have the best agreement. This is supported by the findings of this study that experience as an SLP and experience rating voices were not significant predictors of the sequential effect size, indicating that experience is not protective against sequential effects and therefore does not increase a rater's reliability in these circumstances.

CLINICAL AND RESEARCH IMPLICATIONS

In the clinical practice of voice-specialist SLPs, there are few instances where voices are presented close enough together to suggest concern for these sequential effects, as these effects are rendered minimal after more than 10 minutes between judgements (Attali, 2011). Behrman (2005) indicates that 100% of responding voice clinicians in their study complete an auditory perceptual evaluation with every patient during their evaluation, suggesting it is more common to assess voices individually which minimizes the chance of sequential effects. Anecdotally, some voice clinicians occasionally record voice samples and make ratings at the end of the day from all the patients seen that day. In this circumstance, sequential effects may exist should a mild or moderately dysphonic voice follow a set of more severe ones.

A more likely situation for sequential effects to impact clinical practice is training of a new voice clinician. Auditory-perceptual training can take the form of listening to a series of voices and making judgments about them, then comparing ratings with a more senior/experienced clinician. Chan & Yiu (2006) report on the efficacy of two such training programs, both of which rely on sequential presentations of stimuli. This presentation of mixed voices following one another can lead to sequential effects on the part of the trainee, making establishing an internal anchor difficult. There also exist online training programs that follow the same basic format. Due to the methods of this study, we cannot conclude that a random presentation of voices would prevent sequential effects; however, it is recommended to avoid presenting clusters of voices with similar severity levels, as this was shown to produce a contrast bias.

It is much more likely that participants in research studies will experience a sequential presentation of voices, as this is a common method of presentation for studies of voice rating tasks. Although the average 6-point contrast bias found in this study is not clinically meaningful, it is sufficient to cause an effect in research studies. Knowledge of this effect and possible reasons for it should be a consideration in future studies of the CAPE-V and voice rating tasks to avoid drawing inappropriate conclusions.

LIMITATIONS

Some limitations of this project include the sample size, nonuniform distribution of voice samples used, and non-standardized listening conditions. The sample size was chosen because it was realistic for this project's aims; however, a larger sample size would increase the statistical power of the observations made. If more than four test stimuli were used to test this effect, it would increase the validity of the observed effects. It would also be interesting to discover if the sequential effect observed changes in direction or magnitude if the test stimuli were mild or severe rather than moderate in nature to begin with.

The second limitation was the non-normal distribution of voice samples used. They were clustered to the mild end of the CAPE-V rating spectrum, which is a possible reason why there was no sequential effect observed in the mild condition. Although the test stimuli were selected from the middle of the distribution, they were further from the severe condition voices than the mild in overall severity. This distribution was partly due to the availability of prerecorded dysphonic voices in the Stepp Lab for Sensorimotor Rehabilitation Engineering database. In addition, the CAPE-V does not locate the moderate anchor in the middle, i.e., at 50 on the 100-point VAS. Instead, the mild marker is at 12mm, the moderate marker is at 38mm, and the severe at 74mm (ASHA CAPE-V Form. (n.d.)). This causes a moderately dysphonic voice to be rated closer to mild than severe because the difference between the mild and moderate markers is 10 mm smaller than the difference between the severe and moderate markers.

Third, due to the remote nature of this experiment, it was not possible to truly standardize the listening conditions across raters. An effort was made to do so by requiring each rater to be in a quiet environment with a reliable internet connection and to use headphones; however, not all headphones are made equally. The quality of each rater's computer, headphones, and listening environment are all variables that may have influenced their perception and rating of voices. For example, in Pickens et al. (2018), two of the four tested headphone brands did not have adequate reliability for hearing tests when compared to results of a hearing screening performed in an audiologist's office. This indicates that headphone quality is varied, and therefore an uncontrolled variable in this study. However, the benefit of conducting the project remotely was the ability to gather participants from a wide geographic range which leads to a more diverse sample and allows conclusions to be drawn about a wider array of voice raters.

Finally, the design of this study relied on clustered presentations of voices followed by a test stimulus on which the sequential effect could be observed. This does not allow for conclusions to be drawn about potential sequential effects resulting from a random presentation of voices, or to determine the minimum exposure needed to create a condition for sequential effects. It could be that one severe voice is sufficient to cause sequential effects in judgement; future studies of this phenomena should consider this question as well.

FUTURE DIRECTIONS

The purpose of this project was to establish if this particular type of perceptual bias—sequential effects—exists in ratings of voice quality. Since this effect was established to exist, it opens the door for future studies to determine when and why this effect exists, and what can be done to prevent it. Some possible directions for this exploration include examining sequential effects on mild and severe voices rather than only moderate ones. Although this study found a contrast bias in the severe condition only, it would be interesting to see if test stimuli further from the mild condition voices elicit a contrast bias in that condition as well. Another factor to examine is how much time between ratings is required to minimize the sequential effect, or what the minimum condition that causes sequential effects would be.

Additionally, a more in-depth analysis of protective factors is indicated, as the scope of this study in that regard was minimal. This could include factors such as the intra-rater reliability of each rater and the amount of time taken to make each rating. A more sensitive measure of rater confidence is also indicated as an area to delve deeper, as that was shown to have a small effect but was not within the scope of this project to explore further.

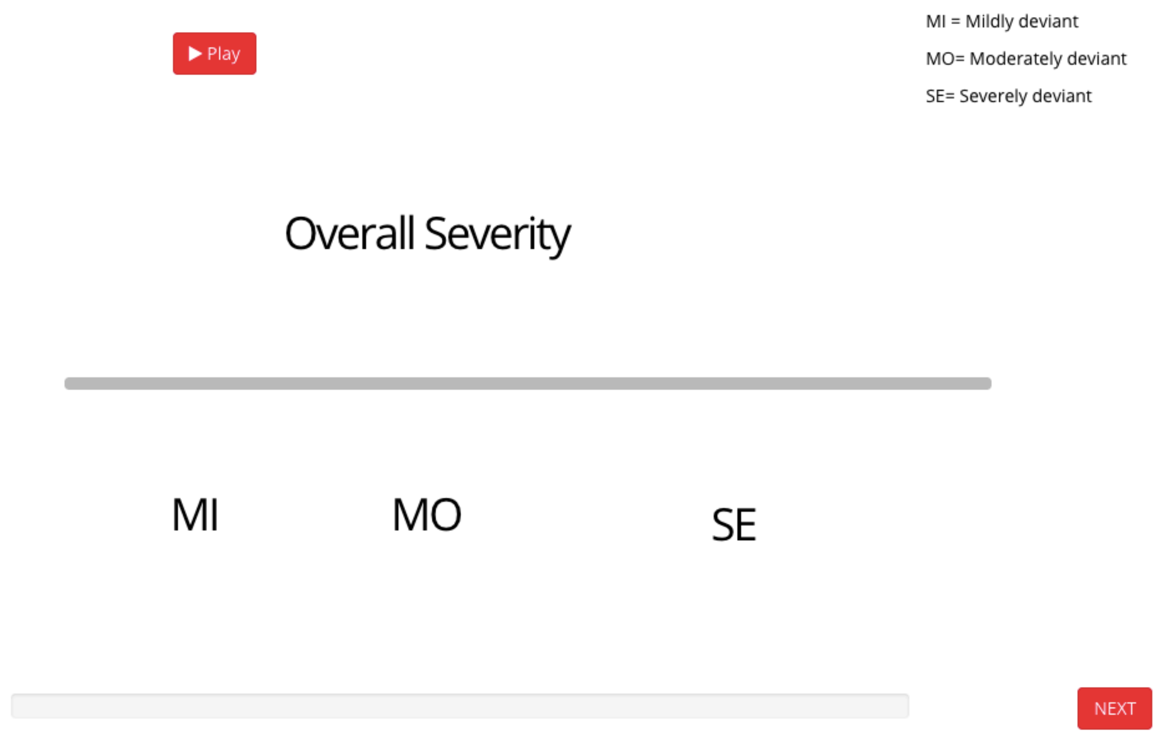
Although this project focused only on the overall severity of dysphonia rating, sequential effects could be at play in ratings of other vocal auditory percepts. Due to the methods of this study, it cannot be concluded that sequential effects will occur in ratings of breathiness, roughness, or strain. However, it would be interesting to assess if they do occur and how they differ. These percepts will be more challenging to study because of lower clinician reliabilities in ratings of breathiness, roughness, and strain (Karnell et al., 2007; DeBodt et al., 1997; Shrivastav, 2006.)

Additionally, although the CAPE-V was the choice of instrument in this study, the GRBAS is also frequently used in the voice clinician community. It uses a 4-point scale of normal (0), mild (1), moderate (2), or severe (3) to rate voices on five aspects of voice quality. This study's methods could be replicated using the GRBAS to compare any sequential effects that exist using that rating scale. Since the GRBAS uses a 4-point scale, it is less likely to show sequential effects because a moderate voice's rating would have to cross a categorical boundary following mild or severe voices to show a sequential effect. However, if a sequential effect was seen, it would be more clinically significant than the effect shown in this study because the gross resolution of the scale would cause a voice to change categories entirely.

CONCLUSION

This study examined sequential effects in overall severity of dysphonia ratings using the CAPE-V for raters with and without experience rating voices. A contrast bias was found for all raters in the severe condition, meaning moderate voices were rated on average six points less severe when following a group of five severe voices. However, raters without experience rating voices showed a larger contrast bias. In addition, confidence in rating voices in general using the CAPE-V increased the size of the contrast bias, while confidence “today” reduced it. These results suggest that while experience rating voices and greater confidence rating voices in a sequential paradigm does reduce the size of the contrast bias, the sequential effect still exists. Results from this study indicate caution in using a clustered presentation of voices for ratings as they are susceptible to this effect, and consideration of sequential effects should be incorporated into future studies of the CAPE-V and voice ratings.

APPENDIX A: Participant View



APPENDIX B: Post-experiment survey

How many years have you been a practicing SLP? (including CF)

0

0 — 40

How many years have you been using the CAPE-V clinically?

0

0 — 20

Please rate your confidence level using the CAPE-V clinically.

Not confident — **Very confident**

Please rate your confidence level in your ratings today.

Not confident — **Very confident**

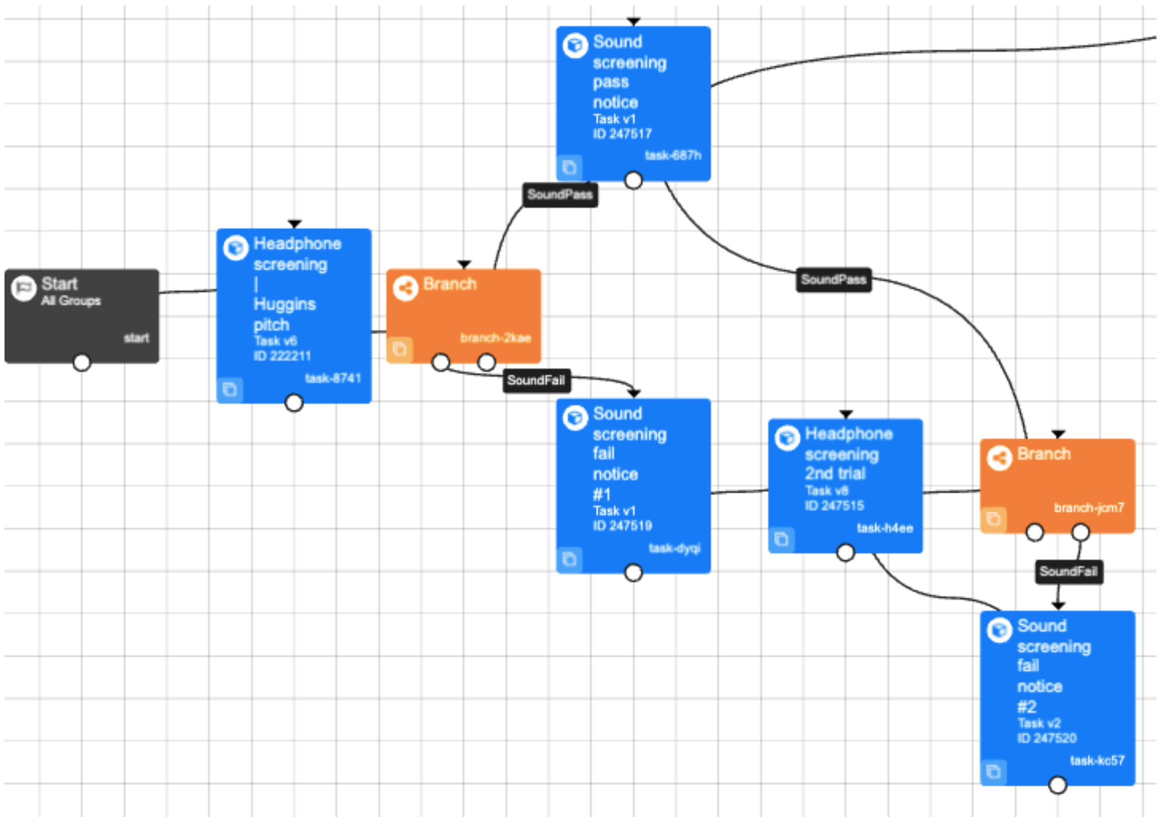
Do you have formal singing voice training above the high school level? If so, how many years?

Thanks for participating! Please return to the Zoom call with the researcher for a debrief and final documentation after clicking "NEXT" below.

Next

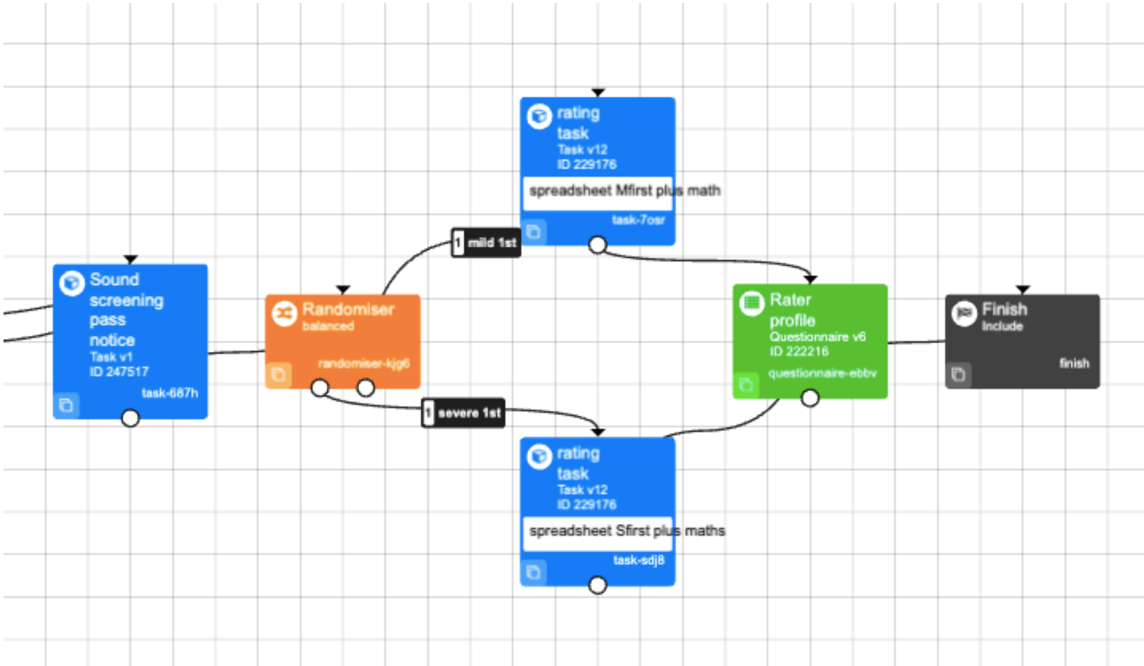
APPENDIX C: Experimental flow diagrams from Gorilla

Figure 3: Headphone screening process



Note: All participants begin at “Start” icon and are directed to the “Headphone screening 1.” If they pass, they receive a pass notice and continue to the experiment. If they fail screening 1, they receive “Sound screening fail notice #1” and are directed to check their audio output and headphone connection. At that point they are directed to “Headphone screening 2nd trial,” after which they either pass and receive a pass notice or fail and receive “sound screening fail notice #2” and are directed to return to the Zoom room for aid from the researcher. A special code is needed to advance to a repeat of “Headphone screening 2nd trial,” and participants are excluded from the experiment if they fail a third time.

Figure 4: Rating task process



Note: After participants receive the “Sound screening pass notice,” the “Randomizer” sends half of the participants from each group to the “mild 1st” branch and the other half to the “severe 1st” branch where the experimental task is located. All participants then complete the “Rater profile Questionnaire” to finish the session.

REFERENCES

Akoglu, Haldun. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91–93. Doi: 10.1016/j.tjem.2018.08.001

Anwyl-Irvine, A.L., Massonié J., Flitton, A., Kirkham, N.Z., & Evershed, J.K. (2019). Gorilla in our midst: an online behavioral experiment builder**.** *Behavior Research Methods,* 52, 388–407. https://doi.org/10.3758/s13428-019-01237-x

ASHA CAPE-V Form. (n.d.). Retrieved April 02, 2021, from https://www.asha.org/form/cape-v/

American Speech-Language-Hearing Association. (n.d.). *Voice Disorders.* (Practice Portal). Retrieved Nov. 3, 2020, from www.asha.org/Practice-Portal/Clinical-Topics/Voice-Disorders/.

Arzounian, D., de Kerangal, M., & de Cheveigné, A. (2017). Sequential dependencies in pitch judgments. *The Journal of the Acoustical Society of America,* 142(5), 3047–3057. DOI: 10.1121/1.5009938

Attali, Y. (2011). Sequential Effects in Essay Ratings. *Educational and Psychological Measurement,* 71(1), 68–79. DOI: 10.1177/0013164410387344

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception & Psychophysics,* 77(4), 1033–1043. https://doi.org/10.3758/s13414-015-0875-8

Behrman A. (2005). Common practices of voice therapists in the evaluation of patients. *Journal of Voice*, 19(3), 454–469.

Best, S. & Fakhry, C. (2011). The prevalence, diagnosis, and management of voice disorders in a National Ambulatory Medical Care Survey (NAMCS) cohort. *The Laryngoscope*, 121(1), 150–157. doi: 10.1002/lary.21169

Boersma, Paul & Weenink, David (2021). Praat: doing phonetics by computer [Computer program]. Version 6.2, retrieved 15 November 2021 from http://www.praat.org/

Chan, K. M. K., & Yiu, E. M. L. (2006). A comparison of two perceptual voice evaluation training programs for naive listeners. *Journal of Voice*, *20*(2), 229+.

https://link.gale.com/apps/doc/A146942704/AONE?u=mlin_b_bumml&sid=bookmark-AONE&xid=7136601e

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, 79,1–39.

Daly, J.A. & Dickson-Markman, F. (1982). Contrast Effects in Evaluating Essays National Council on Measurement in Education. *Journal of Educational Measurement*, 19(4), 309–316.

De Bodt M., Van de Heyning P.H., Wuyts F., & Lambrechts L. (1996). The perceptual evaluation of voice disorders. *Acta Otorhinolaryngologica Belgica* 50, 283–291.

De Bodt, M., Wuyts, F., Van de Heyning, P.H., & Croux, C. (1997). Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*, 11(1), 74–80. https://doi.org/10.1016/S0892-1997(97)80026-4

Dyjas, O., Bausenhart, K.M. & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception & Psychophysics,* 74, 1819–1841 (2012). https://doi.org/10.3758/s13414-012-0362-4

Eadie, T. & Baylor, C. (2006). The Effect of Perceptual Training on Inexperienced Listeners' Judgments of Dysphonic Voice. *Journal of Voice*, 20(4), 527–544. DOI:10.1016/j.jvoice.2005.08.007

Eadie, T., Kapsner, M., Rosenzweig, J., Waugh, P., Hillel, A., & Merati, A. (2010). The role of experience on judgments of dysphonia. *Journal of Voice,* 24(5), 564–573. DOI: 10.1016/j.jvoice.2008.12.005

Eadie, T., Sroka, A., Wright, D., & Merati, A. (2011). Does Knowledge of Medical Diagnosis Bias Auditory-Perceptual Judgments of Dysphonia? *Journal of Voice,* 25(4), 420–429. doi: 10.1016/j.jvoice.2009.12.009.

Eadie, T. L. & Kapsner-Smith, M. (2011). The Effect of Listener Experience and Anchors on Judgments of Dysphonia. *Journal of Speech, Language, and Hearing Research,* 5(2), 430–447. doi: 10.1044/1092-4388(2010/09-0205)

Eargle, J.M. (1994). Frequency and Pitch Relationships. *Electroacoustical Reference Data.* Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-2027-6_142

Fairbanks, G. (1960). *Voice and Articulation Drillbook.* New York, NY. Harper's Press.

Francis, A. L., & Ciocca, V. (2003). Stimulus presentation order and the perception of lexical tones in Cantonese. *The Journal of the Acoustical Society of America*, 114(3), 1611–1621. https://doi.org/10.1121/1.1603231

Gerratt, B.R., Kreiman, J., & Garellek, M. (2016). Comparing measures of voice quality from sustained phonation and continuous speech. *Journal of Speech, Language, and Hearing Research,* 59, 994–1001. doi: 0.1044/2016_JSLHR-S-15-0307

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+. https://link.gale.com/apps/doc/A232946667/AONE?u=mlin_b_bumml&sid=bookmark-AONE&xid=5a2cb438

Hirano M. (1981). *Clinical Examination of Voice*. New York, NY: Springer Verlag; pp. 81– 84.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140, 1–55

Lockwood, P., & Kunda, Z. (1997). Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology*, 73, 91–103.

Karnell, M., Melton, S., Childes, J., Coleman, T., Dailey, S., & Hoffman, H. (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. DOI: 10.1016/j.jvoice.2006.05.001

Kelchner, Lisa N., Brehm, Susan B., Weinrich, B., Middendorf, J., de Alarcon, 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. https://doi.org/10.1016/j.jvoice.2008.09.004

Kempster, G.B., Kistler, D.J., & Hillenbrand, J. (1991). Multidimensional scaling analysis of dysphonia in two speaker groups. *Journal of Speech and Hearing Research,* 34, 534–543.

Kempster, G., Gerratt, B., Verdolini Abbott, K., Barkmeier-Kraemer J., & Hillman, R. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol. *The American Journal of Speech and Language Pathology,* 18(2), 134–142. DOI: 10.1044/1058-0360(2008/08-0017).

Kenrick D.T., & Gutierres S.E. (1980). Contrast Effects and Judgments of Physical Attractiveness: When Beauty Becomes a Social Problem. *Journal of Personality and Social Psychology*, 38(1):131–140

Kidd, G.R., Watson, C.S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America,* 122, 418. doi: 10.1121/1.2743154

King, M. & Lockhead, G. (1981). Response scales and sequential effects in judgment. *Perception & Psychophysics,* 30(6), 599–603.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erma, 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.

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, 2354 (2007); doi: 10.1121/1.2770547

Ma, E.P., & Yiu, E.M. (2006). Mutiparametric evaluation of dysphonic severity. *Journal of Voice*. 20, 380–390.

Martin, L. L., Seta, J. J., & Crelia, R. A. (1990). Assimilation and Contrast as a Function of People's Willingness and Ability to Expend Effort in Forming an Impression. *Journal of Personality and Social Psychology*, 59(1), 27–37. https://doi.org/10.1037/0022-3514.59.1.27

Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Milne, A.E., Bianco, R., Poole, K.C., Zhao, S., Oxenham, A.J., Billig, A.J., Chait, M. (2020) An online headphone screening test based on dichotic pitch. *Behavior Research Methods,* 53, 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com)

Merrill R., Anderson A., & Sloan A. (2011). Quality of life indicators according to voice disorders and voice-related conditions. *The Laryngoscope*, 121(9), 2004–2010. doi: 10.1002/lary.21895

Miyake, K., & Zuckerman, M. (1993). Beyond personality impressions: Effects of physical and vocal attractiveness on false consensus, social comparison, affiliation, and assumed and perceived similarity. *Journal of Personality*, 61,411–437.

Mussweiler, T. (2003). Comparison Processes in Social Judgment: Mechanisms and Consequences. *Psychological Review*, 110(3), 472–489. DOI: 10.1037/0033-295X.110.3.472.

Nemr, K., Simões-Zenari, M., Ferro Cordeiro, G., Tsuji, D., Itar Ogawa, A., Tibério Ubrig, M., & Moreira Menezes, M. H. (2012). GRBAS and Cape-V Scales: High Reliability and Consensus When Applied at Different Times. *Journal of Voice*, 26(6), 812e17–e22. https://doi.org/10.1016/j.jvoice.2012.03.005

Nosanchuk, T. A., & Erickson, B. H. (1985). How high is up? Calibrating social comparison in the real world. *Journal of Personality and Social Psychology*, 48, 624–634.

Oates, J. (2009). Auditory-Perceptual Evaluation of Disordered Voice Quality: Pros, Cons and Future Directions. *Folia Phoniatrica & Logopaedica,* 69, 41–56. DOI: 10.1159/000200768.

Pegors T., Mattar M., Bryan P., & Epstein R. (2015). Simultaneous perceptual and response biases on sequential face attractiveness judgments. *Journal of Experimental Psychology General*, June; 144(3): 664–673. DOI:10.1037/xge0000069.

Pelham, B. W., & Wachsmuth, J. O. (1995). The waxing and waning of the social self: Assimilation and contrast in social comparison. *Journal of Personality and Social Psychology*, 69, 825–838.

Pickens, A., Robertson, L. D., Smith, M. L., Zheng, Q., & Song, S. (2018). Headphone Evaluation for App-Based Automated Mobile Hearing Screening. *International Archives of Otorhinolaryngology*, 22(4), 358–363. https://doi.org/10.1055/s-0037-1607438

Reghunathan, S. & Bryson, P. (2019). Components of Voice Evaluation. *Otolaryngologic Clinics of North America*, 52(4), 589–595. doi: 10.1016/j.otc.2019.03.002

Roy, N., Merrill, R. M., Thibeault, S., Parsa, Rahul A., Gray, S. D., & Smith, E. M. (2004). Prevalence of Voice Disorders in Teachers and the General Population. *Journal of Speech, Language, and Hearing Research*, 47(2), 281–293. https://doi.org/10.1044/1092-4388(2004/023)

Roy N., Merrill R., Gray S., & Smith E. (2005). Voice Disorders in the General Population: Prevalence, Risk Factors, and Occupational Impact. *The Laryngoscope* 115(11), 1988–1995. doi: 10.1097/01.mlg.0000179174.32345.41

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. In H. Martin & A. Tesser (Eds.) *The construction of social judgment* (pp. 217–245). Hillsdale, NJ: Erlbaum.

Sherif, M., & Hovland, C.I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. New Haven, CT: Yale University Press.

Shrivastav, Rahul. (2006). Multidimensional Scaling of Breathy Voice Quality: Individual Differences in Perception. *Journal of Voice,* 20(2), 211–222. https://doi.org/10.1016/j.jvoice.2005.04.005

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. https://doi.org/10.1016/j.jvoice.2009.10.007

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 Archive of Otorhinolaryngology,* 261, 429–434.

Woods, K.J.P., Siegel, M.H., & Traer, J. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics.* 79, 2064–2071. https://doi.org/10.3758/s13414-017-1361-2

Yiu, E.M. (1999). Limitations of perturbation measures in clinical acoustic voice analysis. *Asian Pacific Journal of Speech Language and Hearing*. 4, 155–166

Zhao H., Andersson B., Guo B., & Xin T. (2017). Sequential Effects in Essay Ratings: Evidence of Assimilation Effects Using Cross-Classified Models. *Frontiers in Psychology*, 8(JUN), 933. doi: 10.3389/fpsyg.2017.00933.

Zoom Video Communications, Inc. (2020). *ZOOM cloud meetings* (Version 4.6.9) [Web Version] https://zoom.us

Zraick, R., Kempster, G., Connor, N., Thibeault, S., Klaben, B., Bursac, Z., Thrush, C., & Glaze, L. (2011). Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *American Journal of Speech Language Pathology*, 20(1), 14–22. DOI: 10.1044/1058-0360(2010/09-0105)

CURRICULUM VITAE

[redacted]

[redacted]