AJSLP

Research Article

Reliability and Accuracy of Expert Auditory-Perceptual Evaluation of Voice via Telepractice Platforms

Kimberly L. Dahl,[a] Hasini R. Weerathunge,[a,b] Daniel P. Buckley,[a,c] Anton S. Dolling,[a] Manuel Díaz-Cádiz,[a] Lauren F. Tracy,[c] and Cara E. Stepp[a,c]

Purpose: This study assessed the reliability and accuracy of auditory-perceptual voice evaluations by experienced clinicians via telepractice platforms.
Method: Voice samples from 20 individuals were recorded after transmission via telepractice platforms. Twenty experienced clinicians (10 speech-language pathologists, 10 laryngologists) evaluated the samples for dysphonia percepts (overall severity, roughness, breathiness, and strain) using a modified Consensus Auditory-Perceptual Evaluation of Voice. Reliability was calculated as the mean of squared differences between repeated ratings (intrarater agreement), and between individual and group mean ratings (interrater agreement). Repeated measures analyses of variance were constructed to measure effects of transmission condition (e.g., original recording, WebEx, Zoom), dysphonia percept, and their interaction on intrarater agreement, interrater agreement, and average ratings. Significant effects were evaluated with post hoc Tukey's tests.
Results: There were significant effects of transmission condition, percept, and their interaction on average ratings, and a significant effect of percept on interrater agreement. Post hoc testing revealed statistically, but not clinically, significant differences in average roughness ratings across transmission conditions, and significant differences in interrater agreement for several percepts. Overall severity had the highest agreement and strain had the lowest.
Conclusion: Telepractice transmission does not substantially reduce reliability or accuracy of auditory-perceptual voice evaluations by experienced clinicians.

The delivery of clinical voice care via telepractice has become a topic of increasing interest among laryngologists and speech-language pathologists (SLPs). Telepractice connects clinicians and patients for clinical care administered via telecommunications technology (American Speech-Language-Hearing Association [ASHA], n.d.). This modality allows clinicians to reach patients facing significant barriers to care. Such patients include individuals in remote areas, with conditions requiring specialized care, with significant mobility impairments (Mashima & Doarn, 2008), or with limited English proficiency (Mashima, 2012).

This interest in telepractice became an acute need as the coronavirus pandemic in 2020–2021 placed limitations on in-person practice. An abrupt shift toward telepractice was spurred by general concerns about virus transmission and by specific concerns about elevated risks during voice evaluation and treatment (Cantarella et al., 2020; Doll et al., 2021; Thamboo et al., 2020). This sudden need for changes in the delivery of voice care was supported by a growing body of evidence on the effectiveness of voice treatment via telepractice.

Voice teletherapy has been successfully administered for a variety of voice disorders, including muscle tension dysphonia (MTD; Rangarathnam et al., 2015), vocal fold nodules (Fu et al., 2015; Mashima et al., 2003), laryngeal edema (Mashima et al., 2003), unilateral vocal fold paralysis (Mashima et al., 2003), and Parkinson-associated hypokinetic dysarthria (Constantinescu et al., 2010). Researchers found that treatment outcomes via telepractice were comparable to those achieved via in-person treatment. However, in most of these studies, pretreatment voice evaluations were conducted in person, in the clinical research setting, which

[a]Department of Speech, Language and Hearing Sciences, Boston University, MA
[b]Department of Biomedical Engineering, Boston, University, MA
[c]Department of Otolaryngology—Head & Neck Surgery, Boston, University School of Medicine, MA
Correspondence to Kimberly L. Dahl: dahl@bu.edu

Editor-in-Chief: Katherine C. Hustad
Editor: Rita R. Patel
Received March 24, 2021
Revision received May 23, 2021
Accepted May 24, 2021
https://doi.org/10.1044/2021_AJSLP-21-00091

Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

typically entails a sound-treated environment and specialized equipment. Thus, while the evidence suggests that telepractice is an effective delivery method for many behavioral voice treatments, there is little evidence that voice evaluations—on which accurate diagnosis and treatment monitoring depend—can be effectively carried out via telepractice platforms.

A comprehensive voice evaluation includes a case history, patient-reported outcome measures, laryngeal imaging, acoustic and aerodynamic measures, and auditory-perceptual assessment (Patel et al., 2018; Roy et al., 2013). These evaluation elements vary in the degree to which they can be successfully administered via telepractice. Case histories and patient-reported measures can be easily collected via virtual platforms, while laryngeal visualization cannot. In fact, ASHA has issued specific guidance to allow for the provision of voice therapy without laryngeal visualization in an effort to balance professional ethics with the need to protect patients and clinicians from coronavirus exposure (ASHA, 2020). These guidelines state that recommendations for voice therapy may be based solely on medical history and a limited voice evaluation consisting of auditory-perceptual assessment and acoustic analysis, if available. Although acoustic measures can be collected remotely, a recent study has shown that accuracy of many of these measures is significantly reduced when based on signals transmitted via telepractice platforms (Weerathunge et al., 2021). Auditory-perceptual evaluations can be completed through telepractice platforms, but there is little evidence of their reliability and accuracy.

Only two studies, to our knowledge, have investigated voice evaluations conducted via telepractice. One early study (Duffy et al., 1997) involved the evaluation of patients with voice, speech, language, cognitive, and swallowing complaints. The evaluations were conducted via satellite projection of audiovisual signals to television monitors with a clinician-controlled camera. The effectiveness of the telepractice evaluations was determined by the certainty with which a diagnosis was reached. Diagnostic certainty was shown to be high among participants with voice complaints (81%). However, the technology used in this study is no longer applicable to telepractice today, and the study did not establish clinicians' ability to determine disorder severity along with diagnosis. Finally, the contribution of auditory-perceptual evaluations to the diagnostic process was not specifically assessed.

In another study (Constantinescu et al., 2010), researchers compared auditory-perceptual and acoustic measures of voice of individuals with Parkinson's disease (PD) that were simultaneously collected via a virtual platform and in-person. Perceptual evaluations were completed using 5-point Likert scales for breathiness, roughness, strain, vocal tremor, pitch, loudness, and phonation breaks. Interrater reliability was high when measured by percent close agreement but failed on most voice parameters to meet the authors' clinical criterion for acceptability using a quadratic-weighted Kappa statistic.

The findings of Constantinescu et al. (2010) are difficult to interpret in the present environment of voice telepractice for several reasons. The sample of speakers was limited to individuals with PD, and the sample of raters was limited to SLPs. Thus, neither the broad range of voice disorders seen in a typical voice clinic nor the essential role laryngologists play in diagnosis of voice disorders were captured in the study. Additionally, the virtual platform was custom-built for the study, the set-up involved special equipment such as a clinician-controlled web camera, and the participant attended the session at the research site. Telepractice today is largely carried out via third-party, web-based platforms with patients attending sessions on personal devices from home (Grillo, 2017; Weidner & Lowman, 2020). Under pandemic restrictions, clinicians may also lead sessions from home. Furthermore, the study used nonstandardized rating scales, whereas in clinical practice, standardized instruments such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster et al., 2009) are commonly used. Finally, this study did not include an overall measure of dysphonia, although past research has shown ratings of overall severity to have greater reliability than ratings of individual voice parameters (Helou et al., 2010; Zraick et al., 2011).

There is reason to believe that auditory-perceptual evaluations of voice conducted via telepractice may not be equivalent to evaluations conducted in person. Telepractice conditions are apt to introduce noise to the signal received by the clinician. This noise may originate from the patient's home environment or from the online transmission of the signal. Videoconferencing platforms also use dynamic signal enhancements, such as noise suppression, high-pass filters, and automatic gain control (Cisco Webex, 2021; Zoom Video Communications, Inc., 2021). Though these enhancements aim to improve sound quality for platform users, they may undermine virtual voice assessment by distorting the signal that serves as the basis of the auditory-perceptual evaluation.

Thus, many questions remain about the reliability and accuracy of virtual voice assessment. Although current conditions of telepractice may not continue indefinitely, voice care via telepractice was growing prior to the coronavirus pandemic and is likely to remain a part of the voice clinician's caseload. A better understanding of any advantages or limitations of voice evaluations conducted via telepractice is needed.

A specific assessment of the telepractice tools widely used in voice care is of particular interest. Such tools include the videoconferencing software programs Cisco Webex and Zoom, which are the platforms most commonly used by voice clinicians providing treatment via telepractice (Grillo, 2017). Cisco Webex (Cisco Systems) and Zoom (Zoom Video Communications) are compliant with the Health Insurance Portability and Accountability Act. They offer basic functionalities necessary for effective provision of voice care via telepractice, such as screen sharing, encryption, deletion of transmitted data, and automatic audio/video adjustments for bandwidth optimization (Grillo, 2019).

The purpose of this study was to assess the reliability and accuracy of expert auditory-perceptual evaluations

of voice via these popular telepractice platforms. More specifically, we sought to assess the reliability and accuracy of CAPE-V ratings by SLPs and laryngologists based on voices transmitted via Cisco WebEx, Zoom with default audio enhancements, and Zoom without audio enhancements. We hypothesized that auditory-perceptual evaluations of voice via these telepractice platforms would show lower reliability than in-person evaluations of voice. We also expected that auditory-perceptual evaluations via telepractice platforms would be less accurate, as indicated by a significant difference in mean ratings in telepractice conditions. Finally, we hypothesized that reliability for dysphonia percepts that tend to show lower reliability in typical circumstances (e.g., strain, roughness; Helou et al., 2010; Zraick et al., 2011) would be particularly poor when assessed via telepractice platforms.

Method

Experienced Raters

Experienced raters were 10 laryngologists and 10 SLPs. Raters had at least 3 years experience in the clinical evaluation of voice (M = 10.7 years, SD = 7.7 years, range: 3–20 years). Individuals with voice disorders comprised at least 33% of each rater's caseload. No rater reported any hearing, voice, speech, or language disorder. All raters completed written consent in compliance with the Boston University Institutional Review Board.

Speakers and Speech Stimuli

Speech stimuli were drawn from an existing database of recordings at Boston University. These included recordings of 20 speakers (nine females, 11 males[1]; age: M = 59.3 years, SD = 18.4 years, range: 19–82 years) who had been diagnosed with a voice disorder by a laryngologist or neurologist. These speakers constitute a subset of the participant sample in a recent study by Weerathunge et al. (2021), which assessed the accuracy of acoustic analysis via telepractice platforms. Stimuli included the second of three productions of the sustained vowels /a, i/ (3–5 s) and the second and third sentences of the Rainbow Passage (Fairbanks, 1960). Before and after each vowel production and running speech segment was a 1-s period of silence, which allowed the listener to hear the level and quality of the background noise. All recordings were collected in a sound-treated booth at Boston University using a head-mounted microphone (Shure WH20) with a sampling rate of 44.1 kHz.

A variety of voice disorders commonly evaluated by laryngologists and SLPs were represented in the study sample. Diagnoses included MTD (n = 6), Parkinson-associated dysphonia (n = 6), laryngeal dystonia (n = 4), vocal fold polyps (n = 2), vocal fold nodules (n = 1), and unilateral vocal fold paralysis (n = 1). Blind evaluations of the voice samples were completed by a voice-specializing SLP to ensure an appropriate range of disorder severity. The SLP rated each voice for overall severity using the 100-mm visual analog scale (VAS) of the CAPE-V. Overall severity ranged from 0 to 64.4 (M = 27.2, SD = 17.5), corresponding to a range of typical voice to moderate-to-severe dysphonia. All speakers completed written consent in compliance with the Boston University Institutional Review Board.

Stimuli Preparation

Speech stimuli were transmitted through Cisco Webex and Zoom. Zoom allows users to disable audio enhancement settings to use "original sound," and so speech stimuli were transmitted via Zoom under two conditions—with default audio enhancements and without audio enhancements. Thus, a total of four transmission conditions were included in the study—original recording, Cisco Webex, Zoom with enhancements, and Zoom without enhancements.

The 20 voice recordings described above were transmitted and received by staff of a previous study (Weerathunge et al., 2021) through each of the selected platforms on personal computers. These study staff worked in pairs at their respective homes to prepare the stimuli used in this study; one staff member filled the role of the *transmitter* and the other of the *receiver*. Transmitters played the sound files of the original voice recordings from a handheld recording device (LS-10 Linear PCM Recorder; Olympus Corporation) through an external speaker (soundcore+motion A3116011) placed 58 cm from their computer microphone. This distance was determined through pilot testing and was the average distance between the mouths of 15 speakers and their computer microphones when seated comfortably during a videoconference call. The transmitter's setup thus approximated the typical setup of a patient attending a telepractice session while seated at home in front of their personal computer. Receivers recorded the transmitted sound files directly into Praat (Boersma & Weenink, 2015) on their personal computer at home at a sampling frequency of 44.1 kHz.

Settings on the transmitter's external speaker were chosen to prevent sound distortion and to approximate a sound pressure level comparable to that of typical speech in a quiet room (i.e., 80 dB SPL). Thus, the volume of the external speaker was set to 50% and sound equalization set at 0 dB amplification from 80 Hz to 12 kHz. The volume was set at 100% for the handheld recorder. The final set of stimuli included 80 recordings—the original 20, plus those same recordings transmitted through each of the three platforms.

Listening Procedure

The listening procedure was conducted via Gorilla Experiment Builder (http://www.gorilla.sc), an online research platform. Although Gorilla allows for unsupervised administration of web-based experiments, all experimental sessions for this study were supervised remotely by videoconference. Raters were informed during recruitment that they would participate in a study on telepractice, but were otherwise blinded to the study purpose, including which telepractice platforms were being assessed.

[1]Gender information was not collected.

Raters completed the listening task at an off-site location using a device and headphones of their choosing. Raters first completed a volume-setting task in which they played a 4-s clip of white noise and adjusted the volume on their device to a comfortable level. They then completed an antiphase tone-discrimination task drawn from the Gorilla open materials repository (Milne et al., 2020) to ensure that the participant used headphones to complete the experiment (Woods et al., 2017). During the task, three 200-Hz sinusoidal tones of varying loudness levels were presented to the participant, who was instructed to identify the quietest of the three tones. Raters were required to make three correct identifications to proceed with the experiment.

Raters were informed that they would listen to voice samples and evaluate each sample using a version of the CAPE-V that was modified for online administration. All stimuli, including the original recordings and those recorded after transmission via telepractice platforms, were then presented to each rater in a randomized order, with 20% of samples from each condition repeated at the end of the listening task to measure intrarater reliability. Raters evaluated each voice based on sustained /a, i/ and an excerpt of the Rainbow Passage (Fairbanks, 1960), as described in the Speakers and Speech Stimuli section.

Participants recorded their auditory-perceptual evaluations on a modified CAPE-V rating form, constructed in Gorilla (see Figure 1). The modified CAPE-V included ratings of overall severity, roughness, breathiness, and strain marked on a VAS ranging from 0 to 100. The full set of dysphonia features on the standard CAPE-V was limited to these four percepts to minimize time commitment of participation and facilitate recruitment of practicing SLPs and laryngologists. The Gorilla interface presented the VAS as a continuous input scale, but output was quantized to whole numbers. Each VAS included anchors for mild at 10, moderate at 35, and severe at 72, each corresponding to the anchor positions on the current version of the printable CAPE-V form (http://asha.org/form/cape-v/). Consistency of each parameter was indicated by marking one of two radio buttons labeled "Consistent" and "Intermittent" and positioned to the right of each VAS. Raters were allowed to play each voice recording up to 2 times to complete their evaluation. A second presentation was allowed to mitigate potential disruptions in the rater's home environment that could cause occasional inattention during the first iteration of the recording. However, unlimited presentations were not allowed, to more closely approximate a typical, busy clinical practice. The listening task took approximately 1 hr (M = 62 min, SD = 14.4 min, range: 44–103 min).

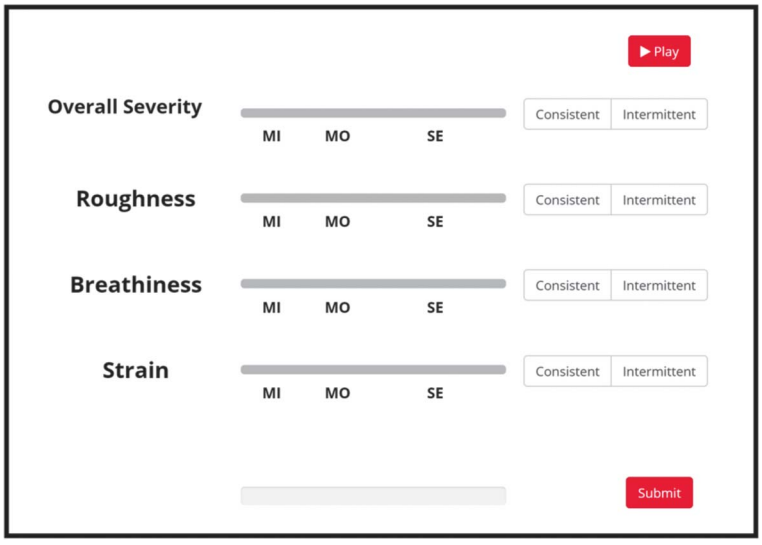
Analysis

Intrarater and Interrater Agreement

Both intrarater and interrater agreement were assessed using variability scores (see Chan & Yiu, 2002; Eadie & Kapsner-Smith, 2011). These variability scores—the square of the difference between ratings—demonstrate the degree to which an individual rating or rater differs from other ratings or raters. Higher scores indicate greater variability and thus less agreement. The use of variability scores offers two benefits over other analyses of reliability or agreement. First, it quantifies individual variability, which can be hidden by measures that assess group-level reliability (Kreiman et al., 1993). Second, as a measure of agreement, rather than reliability, it offers greater clinical applicability (Kreiman et al., 1993).

Intrarater agreement was assessed by calculating a variability score based on ratings of repeated stimuli. The variability score was the square of the difference between

Figure 1. Electronic version of the modified Consensus Auditory-Perceptual Evaluation of Voice rating form, including visual analog scales for ratings of overall severity, roughness, breathiness, and strain; and radio buttons for indicating the consistency ("Consistent" or "Intermittent") of each characteristic.



the first and second ratings of repeated samples. There were four such pairs of repeated ratings per transmission condition, and the variability scores for these four pairs were averaged to generate a mean intrarater variability score for each rater and transmission condition. Mean intrarater variability scores were used in the statistical analysis described below. To illustrate differences between original recordings and telepractice conditions, the mean intrarater variability scores for each condition were normalized to variability scores in the original condition.

Interrater variability scores were calculated as the square of the difference between the individual rating of a sample and the mean rating of the entire group of raters for that sample. Again, the interrater variability scores were averaged for each rater and transmission condition, and mean scores were used in statistical analyses. Interrater variability scores were normalized to the original condition to illustrate differences in agreement between original and telepractice conditions.

CAPE-V Ratings

A total of 320 CAPE-V ratings were collected for each speaker—20 ratings for each of the four dysphonia percepts in each of the four transmission conditions. These individual ratings were used in the statistical analysis described below. For illustration purposes, CAPE-V ratings were also averaged across all listeners to calculate a mean rating for each speaker in each transmission condition and for each percept. These ratings where then normalized to each speaker's mean rating in the original condition.

Statistical Analysis

Three repeated-measures analyses of variance (ANOVAs) were constructed to measure the effects of transmission condition (original, Cisco Webex, Zoom with enhancements, Zoom without enhancements), dysphonia percept (overall severity, roughness, breathiness, strain), and their interaction on intrarater agreement, interrater agreement, and mean CAPE-V rating. Rater was entered as a random variable for ANOVAs assessing agreement, and both rater and speaker for the ANOVA assessing CAPE-V ratings. Effect sizes for each factor in the ANOVAs were calculated as squared partial curvilinear correlations (η_p^2). Effect sizes of ~0.01 were considered small, ~0.9 medium, and > 0.25 large (Witte & Witte, 2009). Significant effects were evaluated with post hoc Tukey's test of multiple comparisons. Effect sizes (Cohen's *d*) were calculated to measure the magnitude of differences and were designated as small (0.25), medium (0.55), or large (> 0.93), per updated recommendations specific to the field of speech, language, and hearing sciences (Gaeta & Brydges, 2020). Statistical analysis was completed in Minitab (version 19.2020.1), with statistical significance set at $\alpha = 0.05$.

Results

Intrarater and Interrater Agreement

Mean variability scores reflecting intrarater and interrater agreement for each transmission condition and dysphonia percept are listed in Tables 1 and 2, respectively. Figure 2 shows intrarater and interrater agreement for each transmission condition, normalized to the variability scores for the original in-person voice recording.

CAPE-V Ratings

Normalized mean CAPE-V ratings are plotted in Figure 3. The magnitude of changes in CAPE-V ratings across transmission conditions was small, ranging from a decrease of 0.2 units for strain ratings in the Zoom without enhancements condition to an increase of 5.0 units for roughness ratings in the Cisco Webex condition.

Statistical Results

There was no significant effect of transmission condition on intrarater or interrater agreement nor a significant interaction between transmission condition and percept. There was a small but significant effect of percept, $F(3, 285) = 11.24$, $p < .05$, $\eta_p^2 = 0.11$, on interrater agreement only. See Tables 3 and 4 for complete ANOVA results.

Post hoc testing with Tukey's multiple comparisons procedure revealed significant differences in interrater agreement for several percepts, all with large effect sizes. There was significantly higher interrater agreement for overall severity (156.57) than for roughness (199.46, $d = 1.24$, $p_{adj} < .05$), breathiness (202.69, $d = 2.54$, $p_{adj} < .05$), and strain (236.12, $d = 3.93$, $p_{adj} < .05$). There was significantly lower interrater agreement for strain than for roughness ($d = -0.98$, $p_{adj} < .05$). There were no significant differences in interrater agreement between roughness and breathiness nor between strain and breathiness (both $p_{adj} \geq .05$).

Table 1. Mean intrarater agreement, calculated as variability score, of Consensus Auditory-Perceptual Evaluation of Voice ratings for each transmission condition and dysphonia percept from a sample of experienced voice clinicians ($n = 20$).

Transmission condition	Overall severity (*SD*)	Roughness (*SD*)	Breathiness (*SD*)	Strain (*SD*)
Original	144.0 (198.7)	209.4 (235.9)	235.7 (225.6)	169.5 (179.5)
Cisco Webex	147.9 (135.4)	193.1 (241.0)	113.4 (141.9)	154.5 (128.5)
Zoom (default)[a]	124.4 (140.0)	114.2 (136.1)	97.6 (96.1)	218.6 (309.7)
Zoom (raw)[b]	83.8 (86.2)	121.6 (115.0)	147.2 (139.1)	156.5 (206.2)

[a]With default audio enhancements. [b]Without audio enhancements.

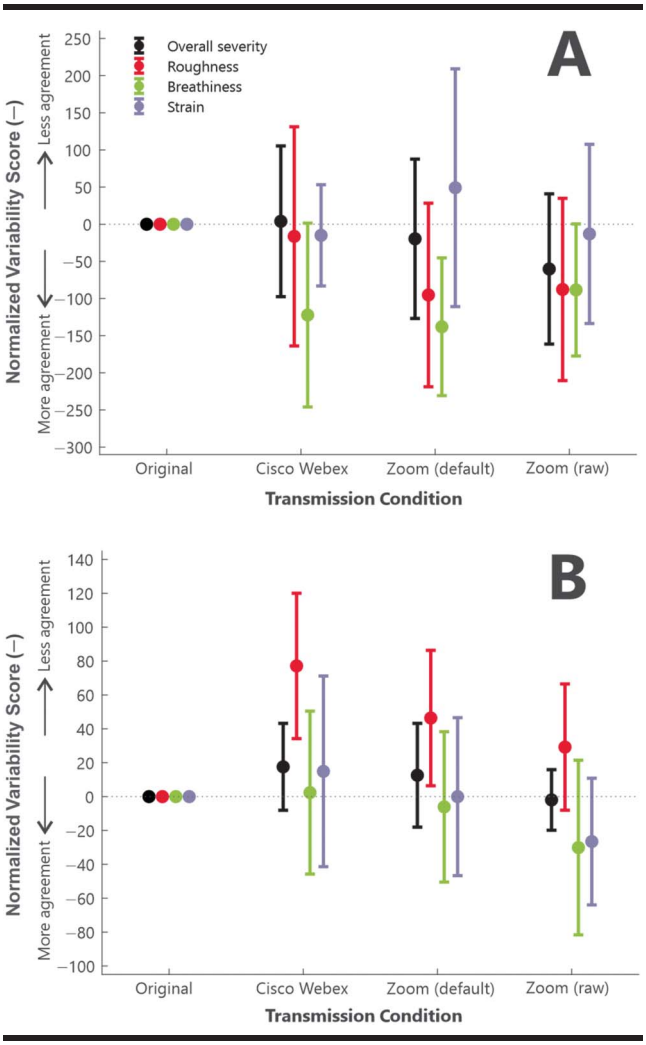
Table 2. Mean interrater agreement, calculated as variability score, of Consensus Auditory-Perceptual Evaluation of Voice ratings for each transmission condition and dysphonia percept from a sample of experienced voice clinicians (n = 20).

Transmission condition	Overall severity (*SD*)	Roughness (*SD*)	Breathiness (*SD*)	Strain (*SD*)
Original	149.5 (92.8)	161.3 (110.5)	211.1 (151.8)	239.0 (152.7)
Cisco Webex	167.1 (101.2)	238.4 (104.5)	213.5 (118.5)	253.9 (130.5)
Zoom (default)[a]	162.1 (104.8)	207.7 (112.2)	205.1 (125.7)	239.0 (129.6)
Zoom (raw)[b]	147.6 (83.8)	190.5 (107.1)	181.1 (93.5)	212.5 (102.2)

[a]With default audio enhancements. [b]Without audio enhancements.

There was a significant effect of condition, percept, and their interaction on mean ratings. See Table 5 for complete ANOVA results. Post hoc Tukey's tests revealed significant differences in mean CAPE-V ratings for two transmission condition and percept interactions, with small effect sizes. Mean roughness ratings were significantly lower in the original condition (23.56 units) than in Cisco Webex (28.58 units, $d = 0.21$, $p_{adj} < .05$) and in Zoom without enhancements (27.99 units, $d = 0.19$, $p_{adj} < .05$).

Figure 2. Variability scores across transmission conditions, normalized to scores in the original recording condition. Panel A shows intrarater agreement. Panel B shows interrater agreement. Zoom (default) is Zoom with default audio enhancements. Zoom (raw) is Zoom without audio enhancements. Error bars are 95% confidence intervals.



Discussion

Reliability of Telepractice Voice Evaluations

The results of this study are encouraging. They show that the reliability of auditory-perceptual evaluations of voice conducted via telepractice platforms by experienced clinicians is comparable to that of in-person auditory-perceptual evaluations. Our reliability analysis shows interrater agreement to be comparable to or slightly lower than that found in previous studies that used the variability score approach taken here (Eadie & Kapsner-Smith, 2011; Helou et al., 2010). Direct comparison with these studies, however, is difficult given methodological differences related to listener training, listener blinding, sample size, and rating scales. Specifically, this study included 20 partially blinded expert listeners who completed evaluations with the CAPE-V. No listener training, which is known to affect reliability

Figure 3. Mean CAPE-V ratings across transmission conditions, normalized to ratings in the original recording condition. Zoom (default) is Zoom with default audio enhancements. Zoom (raw) is Zoom without audio enhancements. Error bars are 95% confidence intervals.

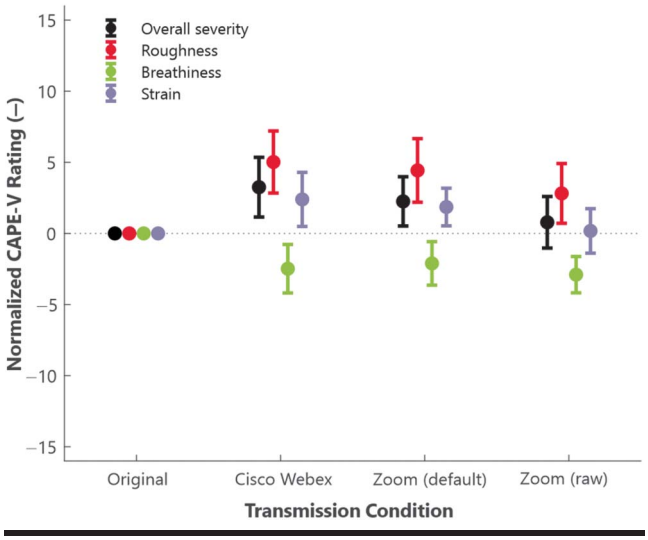


Table 3. Results of repeated-measures analysis of variance for intrarater agreement.

Effect	*df*	*F*	*p*	η_p^2
Transmission condition	3	2.01	.112	NS
Percept	3	1.20	.309	NS
Condition × Percept	9	1.14	.334	NS

Note. NS = not significant.

Table 5. Results of repeated-measures analysis of variance for Consensus Auditory-Perceptual Evaluation of Voice ratings.

Effect	*df*	*F*	*p*	η_p^2	Effect size
Transmission condition	3	8.15	< .001*	.00	Small
Percept	3	234.90	< .001*	.10	Medium
Condition × Percept	9	4.13	< .001*	.01	Small

*Significant at $p < .05$.

(Chan & Yiu, 2002), was provided. In contrast, Helou et al. (2010) enrolled 10 experienced listeners, some of whom were not blinded to the study purpose and were given 10–15 min of CAPE-V training. Eadie and Kapsner-Smith (2011) enrolled 10 experienced listeners who were not offered specific training and who completed evaluations on a VAS scale that differed from the published CAPE-V. These methodological differences may account for the slight differences in variability scores noted in this study.

Nevertheless, in this study, neither intrarater nor interrater agreement for four dysphonia percepts (overall severity, roughness, breathiness, and strain) suffered as an effect of voice samples being transmitted via two popular telepractice platforms—Cisco Webex and Zoom (with and without default audio enhancements). Many of the key elements of a comprehensive voice evaluation cannot be effectively conducted remotely, so the finding that reliability of auditory-perceptual evaluations is maintained in telepractice conditions bodes well for meeting both acute and long-term demands for virtual voice care.

We did find a significant effect of dysphonia percept on interrater reliability, as hypothesized. Interrater agreement was highest for overall severity and lowest for strain. This finding is consistent with past research. Strain is typically found to demonstrate the lowest levels of agreement across listeners (Helou et al., 2010; Kelchner et al., 2010; Zraick et al., 2011), and overall severity often demonstrates higher agreement than individual percepts (Helou et al., 2010; Zraick et al., 2011). This pattern held true across transmission conditions in this study.

Accuracy of Telepractice Voice Evaluations

Rater reliability does not fully address the question of whether a voice evaluation can be effectively completed via telepractice. Raters can, of course, consistently agree on a mischaracterization of a voice. The mean CAPE-V ratings across transmission conditions (see Figure 3), therefore, provide important context for interpreting our reliability findings and drawing conclusions about the accuracy of auditory-perceptual voice evaluations. Although there was a statistically significant effect of transmission condition on CAPE-V ratings, the mean ratings differed little across transmission conditions. Mean differences between telepractice conditions and the original condition ranged from 0.2 to 5.0 units on the 100-unit CAPE-V scale. Even where statistically significant differences in roughness ratings were identified, they were below a clinically significant difference as defined by analogy to the common 7-point equal appearing interval scale, which would require a difference of at least 7.14 units (Eadie & Kapsner-Smith, 2011).

This finding eliminates the possibility that reliability held steady across transmission conditions because of fundamental differences in how the voices were perceived in the different telepractice platforms. It was not the case that the signals were altered in such a way that raters no longer perceived dysphonic characteristics or perceived all voices as being more severely dysphonic. If, for example, a telepractice platform changed the audio signal such that breathiness was rendered imperceptible, we would expect both consistent agreement within and between raters and very low breathiness ratings. Only the former was true of our results, across all dysphonia percepts. Thus, raters consistently identified both the presence and severity of each dysphonia percept across all transmission conditions.

These results are somewhat surprising. Recent work has shown that the acoustics of voice signals are significantly altered when transmitted via telepractice platforms, including those platforms assessed here (Weerathunge et al., 2021). Weerathunge et al. (2021) found that measures sensitive to dysphonia, such as smoothed cepstral peak prominence, low-to-high spectral energy ratio (LHR), and harmonic-to-noise ratio, differed significantly with large effect sizes during telepractice transmission. The magnitude of these changes was near or exceeding values that would distinguish between dysphonic and typophonic voices (e.g., Lowell et al., 2012; Sauder et al., 2017).

Acoustic differences across telepractice conditions were indeed present in the samples evaluated by the raters in the current study and were maintained in transmission through the Gorilla platform. Figure 4 illustrates these differences, reflected in the frequency spectra of two voice samples of a single speaker—the original recording and the

Table 4. Results of repeated-measures analysis of variance for interrater agreement.

Effect	*df*	*F*	*p*	η_p^2	Effect size
Transmission condition	3	2.55	.056	NS	—
Percept	3	11.24	< .001*	.11	Medium
Condition × Percept	9	0.60	.796	NS	—

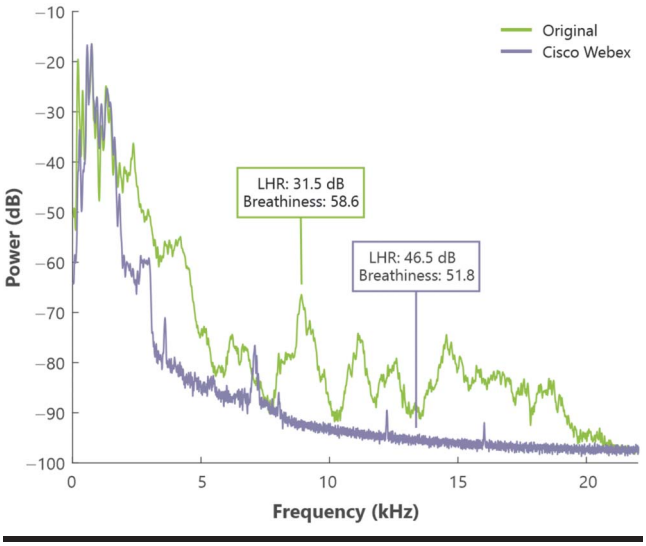
Note. Em dashes indicate data not applicable for nonsignificant findings. NS = not significant.

*Significant at $p < .05$.

Cisco Webex recording—recorded directly from the Gorilla transmission. The differences in the signal are evident both on visual inspection of the spectra and when comparing the LHR of each sample (31.5 dB for original, 46.5 dB for Cisco Webex). These LHR differences would have been expected to yield differences in breathiness ratings, given the relationship between these measures (Hillenbrand & Houde, 1996). That was not the case; the difference in breathiness ratings for this speaker in the original recording and the Cisco Webex transmission (6.8 units) was not clinically significant (see Eadie & Kapsner-Smith, 2011). Thus, while the transmission conditions did lead to acoustic differences in the recordings, as seen previously (Weerathunge et al., 2021), these differences did not appear to affect CAPE-V ratings.

One explanation for this lack of an effect of the acoustic differences is the listener's consideration of the acoustic environment in which a voice sample is presented. The stimuli in this study included periods of silence from the original recording that surrounded each speech task in the sample. These periods of silence were subjected to the same acoustic changes induced by the transmission condition as the speech signal itself. Listeners thus had the opportunity to observe the quality of the background noise in which a voice occurred. This may have allowed raters to distinguish between transmission-related noise and dysphonia-related noise in the signal. Research suggesting that representations of human speech in the auditory cortex are noise-invariant (Khalighinejad et al., 2019) provides some support for this interpretation. Importantly, consideration of the acoustic context would also occur in a typical telepractice session. These silent periods were therefore critical for recreating the conditions in which a telepractice voice evaluation actually occurs. Without them, our findings may have been quite different and less applicable to clinical practice.

Figure 4. Frequency spectra of one speaker's sustained /a/ in the original recording and the Cisco Webex transmission. LHR = low-to-high spectral energy ratio. Breathiness is the mean breathiness rating on the Consensus Auditory-Perceptual Evaluation of Voice by all raters.



Clinical Implications

Voice care via telepractice has been growing in recent years. Until the coronavirus pandemic of 2020–21, however, it did so while clinicians maintained the option to conduct voice evaluations in person with standard clinical protocols. The sudden and broad limitations on in-person care during the pandemic created a new environment in which telepractice delivery was required at all stages of care, including assessment. This raised questions about the feasibility, accuracy, and reliability of virtual voice evaluations. The findings of this study show that auditory-perceptual voice assessments conducted via telepractice appear to be as reliable and accurate as in-person evaluations.

The clinical implications extend beyond current pandemic conditions, however. Auditory-perceptual evaluations are not only conducted during pretreatment assessments. Rather, clinicians continually conduct auditory-perceptual assessments throughout the course of treatment to monitor patient progress. Our findings suggest that clinicians can maintain confidence in this crucial clinical tool even when providing care via telepractice beyond the initial assessment.

Limitations

The auditory-perceptual evaluations conducted for this study were based on prerecorded voice samples. This allowed us to transmit identical samples through each telepractice platform. However, evaluations based on prerecorded samples may differ from the real-time evaluations clinicians often conduct. Raters completed their assessments devoid of visual cues, patient history, and voice complaints, and the flexibility to guide the patient through speech tasks that may best elicit perceptual symptoms of a voice disorder. While some of these conditions are known to introduce bias in auditory-perceptual evaluations (e.g., Eadie et al., 2011), they do, nevertheless, represent the typical conditions of such assessments and were not captured here.

The acoustic signal transmitted during the listening task in some ways represented an idealized version of the signal a clinician may receive during an actual telepractice session. The voice samples were transmitted through telepractice platforms and recorded directly into Praat before being presented to the raters. This process may have minimized additional acoustic changes to the signal that may be imposed from the clinician's end in a typical telepractice session. Given the results of this study, in which voice samples collected in this manner were associated with high reliability and accuracy of auditory-perceptual evaluations, clinicians are encouraged to implement this practice during their telepractice evaluation sessions.

The speech stimuli used in the study were drawn from an existing database of voice samples at Boston University. These recordings included sustained vowels, but not the standard CAPE-V sentences. Thus, the speech stimuli in

this study represented a deviation from the published CAPE-V protocol. Given that the stimuli did include speech tasks that are widely used in auditory-perceptual evaluations—sustained vowels and the Rainbow Passage—we do not expect that this deviation from protocol had a substantial effect on our findings.

This study did not assess the effects other relevant factors may have on auditory-perceptual evaluations. Countless combinations of patient and clinician devices and accessories, internal audio settings, Internet speeds, and ambient noise levels may all contribute to the quality of the audio signal received by the clinician. Future work should investigate the optimal settings and equipment needed to maximize faithfulness of the transmitted signal to the original source.

Conclusions

This study investigated the effect of telepractice platforms on the reliability and accuracy of auditory-perceptual evaluations of voice by experienced clinicians. Our results showed that both reliability and accuracy of evaluations conducted via telepractice were comparable to those conducted in person. While some of the recommended components of a comprehensive voice evaluation (Patel et al., 2018; Roy et al., 2013) cannot be accurately or feasibly conducted via telepractice, our study showed that telepractice is an appropriate modality for auditory-perceptual assessment of voice.

Acknowledgments

This work was supported by Grant R01 DC015570 (awarded to C. E. S.) and Grant T32 DC013017 (awarded to C. A. M. and C. E. S.) from the National Institute on Deafness and Other Communication Disorders and Grant UL1 TR001430 (awarded to D. M. C.) from the National Center for Advancing Translational Sciences. The authors thank Roxanne Segina and Nicole Tomassi for their roles in transmitting and receiving voice samples, and all the expert clinicians who provided auditory-perceptual ratings.

References

American Speech-Language-Hearing Association. (2020). *Considerations when providing voice services in the absence of endoscopic evaluation during COVID-19*. https://www.asha.org/SLP/healthcare/Considerations-When-Providing-Voice-Services-in-the-Absence-of-Endoscopic-Evaluation-During-COVID-19/

American Speech-Language-Hearing Association. (n.d.). *Telepractice* [Practice Portal]. http://www.asha.org/Practice-Portal/Professional-Issues/Telepractice

Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer*. http://www.praat.org

Cantarella, G., Barillari, M. R., Lechien, J. R., & Pignataro, L. (2020). The challenge of virtual voice therapy during the COVID-19 pandemic. *Journal of Voice, 35,* 336–337. https://doi.org/10.1016/j.jvoice.2020.06.015

Chan, K. M. K., & Yiu, E. M.-L. (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. https://doi.org/10.1044/1092-4388(2002/009)

Constantinescu, G., Theodoros, D., Russell, T., Ward, E., Wilson, S., & Wootton, R. (2010). Assessing disordered speech and voice in Parkinson's disease: A telerehabilitation application. *International Journal of Language & Communication Disorders, 45*(6), 630–644. https://doi.org/10.3109/13682820903470569

Cisco Webex. (2021). *Video conferencing: Remove background noise during a Webex meeting or event*. https://help.webex.com/en-us/n70a8os/Remove-Background-Noise-During-a-Webex-Meeting-or-Event

Doll, E. J., Braden, M. N., & Thibeault, S. L. (2021). COVID-19 and speech-language pathology clinical practice of voice and upper airway disorders. *American Journal of Speech-Language Pathology, 30*(1), 63–74. https://doi.org/10.1044/2020_AJSLP-20-00228

Duffy, J. R., Werven, G. W., & Aronson, A. E. (1997). Telemedicine and the diagnosis of speech and language disorders. *Mayo Clinic Proceedings, 72*(12), 1116–1122. https://doi.org/10.4065/72.12.1116

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, 54*(2), 430–447. https://doi.org/10.1044/1092-4388(2010/09-0205)

Eadie, T. L., Sroka, A., Wright, D. R., & Merati, A. (2011). Does knowledge of medical diagnosis bias auditory-perceptual Judgments of dysphonia? *Journal of Voice, 25*(4), 420–429. https://doi.org/10.1016/j.jvoice.2009.12.009

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

Fu, S., Theodoros, D. G., & Ward, E. C. (2015). Delivery of intensive voice therapy for vocal fold nodules via telepractice: A pilot feasibility and efficacy study. *Journal of Voice, 29*(6), 696–706. https://doi.org/10.1016/j.jvoice.2014.12.003

Gaeta, L., & Brydges, C. R. (2020). An examination of effect sizes and statistical power in speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research, 63*(5), 1572–1580. https://doi.org/10.1044/2020_JSLHR-19-00299

Grillo, E. U. (2017). Results of a survey offering clinical insights into speech-language pathology telepractice methods. *International Journal of Telerehabilitation, 9*(2), 25–30. https://doi.org/10.5195/ijt.2017.6230

Grillo, E. U. (2019). Building a successful voice telepractice program. *Perspectives of the ASHA Special Interest Groups, 4*(1), 100–110. https://doi.org/10.1044/2018_PERS-SIG3-2018-0014

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*(3), 248–258. https://doi.org/10.1044/1058-0360(2010/09-0012)

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research, 39*(2), 311–321. https://doi.org/10.1044/jshr.3902.311

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. https://doi.org/10.1016/j.jvoice.2008.09.004

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

Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. *Nature Communications, 10*(1), 2509. https://doi.org/10.1038/s41467-019-10611-4

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality. *Journal of Speech and Hearing Research, 36*(1), 21–40. https://doi.org/10.1044/jshr.3601.21

Lowell, S. Y., Kelley, R. T., Awan, S. N., Colton, R. H., & Chan, N. H. (2012). Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality. *Annals of Otology, Rhinology & Laryngology, 121*(8), 539–548. https://doi.org/10.1177/000348941212100808

Mashima, P. A. (2012). Using technology to improve access to health care for culturally and linguistically diverse populations. *SIG 14 Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse (CLD) Populations, 19*(3), 71–76. https://doi.org/10.1044/cds19.3.71

Mashima, P. A., Birkmire-Peters, D. P., Syms, M. J., Holtel, M. R., Burgess, L. P. A., & Peters, L. J. (2003). Telehealth: Voice therapy using telecommunications technology. *American Journal of Speech-Language Pathology, 12*(4), 432–439. https://doi.org/10.1044/1058-0360(2003/089)

Mashima, P. A., & Doarn, C. R. (2008). Overview of telehealth activities in speech-language pathology. *Telemedicine and E-Health, 14*(10), 1101–1117. https://doi.org/10.1089/tmj.2008.0080

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *BioRxiv,* 2020.07.21.214395. https://doi.org/10.1101/2020.07.21.214395

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Švec, J. G., & Hillman, R. (2018). Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *American Journal of Speech-Language Pathology, 27*(3), 887–905. https://doi.org/10.1044/2018_AJSLP-17-0009

Rangarathnam, B., McCullough, G. H., Pickett, H., Zraick, R. I., Tulunay-Ugur, O., & McCullough, K. C. (2015). Telepractice versus in-person delivery of voice therapy for primary muscle tension dysphonia. *American Journal of Speech-Language Pathology, 24*(3), 386–399. https://doi.org/10.1044/2015_AJSLP-14-0017

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology, 22*(2), 212–226. https://doi.org/10.1044/1058-0360(2012/12-0014)

Sauder, C., Bretl, M., & Eadie, T. (2017). Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV). *Journal of Voice, 31*(5), 557–566. https://doi.org/10.1016/j.jvoice.2017.01.006

Thamboo, A., Lea, J., Sommer, D. D., Sowerby, L., Abdalkhani, A., Diamond, C., Ham, J., Heffernan, A., Cai Long, M., Phulka, J., Wu, Y. Q., Yeung, P., & Lammers, M. (2020). Clinical evidence based review and recommendations of aerosol generating medical procedures in otolaryngology – head and neck surgery during the COVID-19 pandemic. *Journal of Otolaryngology - Head & Neck Surgery, 49,* 28. https://doi.org/10.1186/s40463-020-00425-6

Weerathunge, H. R., Segina, R. K., Tracy, L., & Stepp, C. E. (2021). Accuracy of acoustic measures of voice via telepractice videoconferencing platforms. *Journal of Speech, Language, and Hearing Research, 64*(7), 2586–2599. https://doi.org/10.1044/2021_JSLHR-20-00625

Weidner, K., & Lowman, J. (2020). Telepractice for adult speech-language pathology services: A systematic review. *Perspectives of the ASHA Special Interest Groups, 5*(1), 326–338. https://doi.org/10.1044/2019_PERSP-19-00146

Witte, R. S., & Witte, J. S. (2009). *Statistics* (9th ed.). Wiley.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Zoom Video Communications, Inc. (2021). *Zoom rooms audio guidelines.* Zoom Help Center. https://support.zoom.us/hc/en-us/articles/360025379211-Zoom-Rooms-Audio-Guidelines

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*(1), 14–22. https://doi.org/10.1044/1058-0360(2010/09-0105)