JSLHR

Research Note

The Effect of Visual Sort and Rate Versus Visual Analog Scales on the Reliability of Judgments of Dysphonia

Mara R. Kapsner-Smith,[a] Amanda Opuszynski,[a] Cara E. Stepp,[b,c,d] and Tanya L. Eadie[a]

Purpose: The reliability of auditory-perceptual judgments between listeners is a long-standing problem in the assessment of voice disorders. The purpose of this study was to determine whether a relatively novel experimental scaling method, called visual sort and rate (VSR), yielded stronger reliability than the more frequently used method of visual analog scales (VAS) for ratings of overall severity (OS) and breathiness (BR) in speakers with voice disorders.
Method: Fifty speech samples were selected from a database of speakers with voice disorders. Twenty-two inexperienced listeners provided ratings of OS or BR in four rating blocks: VSR-OS, VSR-BR, VAS-OS, and VSR-BR. For the VAS task, listeners rated each speaker for BR or OS using a vertically oriented 100-mm VAS. For the VSR task, stimuli were distributed into sets of samples with a range of speaker severities in each set. Listeners sorted and ranked samples for OS or BR within each set, and final ratings were captured on a vertically oriented 100-mm VAS. Interrater variability, defined as the mean of the squared differences between a listener's ratings and group mean ratings, and intrarater reliability (Pearson *r*) were compared across rating tasks for OS and BR using paired *t* tests.
Results: Results showed that listeners had significantly less interrater variability (better reliability) when using VSR methods compared to VAS for judgments of both OS and BR. Intrarater reliability was high across rating tasks and dimensions; however, ratings of BR were significantly more consistent within individual listeners when using VAS than when using VSR.
Conclusions: VSR is an experimental method that decreases variability of auditory-perceptual judgments between inexperienced listeners when rating speakers with a range of dysphonic severities and disorders. Future research should determine whether a clinically viable tool may be developed based on VSR principles and whether such benefits extend to experienced listeners.

Auditory-perceptual judgments are a fundamental component of voice quality measurement for clinical and research purposes (Kreiman et al., 1993). Although there is little agreement on which single or complex objective measure underlies different voice qualities, such as breathiness (BR), roughness, or strain (Eadie & Doyle, 2005; Latoszek et al., 2018; Maryn et al., 2010), perceptual assessment appears to be able to take into account the multidimensional nature of the voice signal that contributes to overall voice quality (Awan & Lawson, 2009). Poor voice quality is the reason why many patients with voice disorders seek treatment, and improvement in voice quality is considered one of the primary indicators of treatment success.

The majority of research experiments that measure voice quality include listener judgments using a rating scale. One commonly used method includes the use of an *n*-point rating scale (also referred to as an equal-appearing-interval scale), in which the endpoints are fixed and scaling is performed using whole numbers (i.e., any whole number between "1" and "*n*") that imply "equality" between the numeric components on the scale (Stevens, 1975). These scales are easy for listeners to use, but may artificially

[a]Department of Speech & Hearing Sciences, University of Washington, Seattle
[b]Department of Speech, Language & Hearing Sciences, Boston University, MA
[c]Department of Otolaryngology – Head & Neck Surgery, Boston University School of Medicine, MA
[d]Department of Biomedical Engineering, Boston University, MA
Correspondence to Mara R. Kapsner-Smith: mkapsner@uw.edu

Editor-in-Chief: Bharath Chandrasekaran
Editor: Jack J. Jiang
Received October 24, 2020
Revision received January 6, 2021
Accepted January 29, 2021
https://doi.org/10.1044/2021_JSLHR-20-00623

Disclosure: *Cara E. Stepp has received consulting fees from Altec, Inc./Delsys, Inc., companies focused on developing and commercializing technologies related to human movement. Stepp's interests were reviewed and are managed by Boston University in accordance with their conflict of interest policies. All other authors have declared that no competing interests existed at the time of publication.*

inflate levels of chance agreement between listeners and may be less sensitive to detecting changes in patients' voices over time when there are relatively few fixed points on the scale (Karnell et al., 2007). In addition, these scales may be inappropriate for measuring particular voice quality dimensions that are nonlinear because the distances between the numeric components on the scale are, in fact, not equidistant perceptually (Eadie & Doyle, 2002).

As a result of the difficulties with *n*-point scaling, many voice quality experiments use a visual analog scale (VAS). A VAS is typically a 100-mm undifferentiated line with labeled endpoints; in measuring voice quality, one endpoint may be marked "typical/healthy," and the other endpoint is marked as the maximum severity of the dimension under evaluation. A listener indicates the severity of the voice quality for the speaker under evaluation by marking the point on the 100-mm line that corresponds with the perceived severity of that dimension, with a higher value indicating greater severity (i.e., worse voice quality; Kempster et al., 2009). Scores are measured from the endpoint labeled "normal" (0 mm) to the listener's mark (total indicated in mm). While a few studies have shown advantages in listener reliability for *n*-point scales over VASs (Wuyts et al., 1999), more recent studies that include larger numbers of listeners have shown that VASs are at least as reliable within and between listeners as *n*-point scales (Zraick et al., 2011). The increased resolution provided by VASs also has been proposed to offset the biases that listeners may demonstrate in subdividing the lower end of *n*-point scales for voice dimensions, such as overall severity (OS), which are known to vary nonlinearly (Kempster et al., 2009). Finally, VASs have been shown to be more sensitive than *n*-point scales (Karnell et al., 2007; Kreiman et al., 2007; Nemr et al., 2012). However, there is still an ongoing debate as to whether VASs offer ratio- or interval-level data for voice quality dimensions, as for mood and pain (Myles et al., 1999; Price et al., 1983; Zealley & Aitken, 1969).

Despite some of the relative advantages of using a VAS to measure voice quality, reliability of judgments made with these scales may also be problematic, in addition to other limitations related to the task, the rater, and difficulties assessing multidimensional voice signals (Nagle, 2016; Wuyts et al., 1999). Furthermore, most experiments that include either an *n*-point scale or a VAS use an unanchored approach, in which listeners compare the speaker's voice quality to their own internal standards for the voice quality dimension being rated. Internal standards are shaped by experience and memory, and may be unstable and susceptible to change, even within a single listening session (Kreiman et al., 1993). The comparison with internal standards may result in increased variability, particularly between listeners (Kreiman et al., 1993), and thereby pose challenges for both VAS and *n*-point scales.

To address these limitations, several alternative tasks have been suggested (paired comparison, matching with a known stimulus, comparison with anchors, etc.). Such tasks have in common the use of external comparisons, which may improve listener reliability by bypassing idiosyncratic internal perceptual standards, which often vary across listeners. In a traditionally defined psychometric paired comparison task, listeners compare all possible pairs of stimuli in both A–B and B–A orders and judge the extent to which the stimuli are similar or different (Stevens, 1975). In this task, listeners rank the order within the pairs. The main limitation with this approach is that the number of pairs that need to be compared increases nonlinearly with the number of stimuli, which may lead to an overwhelming number of pairs that need to be evaluated within a given study (to N^2, with N being the number of stimuli evaluated within a study). Measures derived from paired comparisons also may be difficult to interpret because they result in ordinal data, although several transformative models have been proposed to derive interval level data (Thurstone, 1994).

A second method that overcomes a listener's reliance on unstable internal referents is the use of anchors (referent samples). In this approach, anchor stimuli are provided that represent positions on the rating scale to illustrate the severity of the dimension being evaluated. These external referents serve as explicit comparisons to which listeners make their judgments of individual speaker stimuli. Typically, anchor stimuli possess archetypical properties relevant for the dimension under study. Anchors have been shown to increase listener reliability in the use of both *n*-point scales and VASs (Awan & Lawson, 2009; Eadie & Kapsner-Smith, 2011); however, it is difficult to obtain naturally dysphonic samples that represent a known severity level for a particular voice quality dimension along any rating scale. Assignment of anchors to fixed scale values may systematically increase or decrease ratings, limiting comparison between ratings made under different anchor conditions (Eadie & Kapsner-Smith, 2011). In addition, most naturally dysphonic samples do not vary as a function of a single voice quality dimension and they cannot easily be validated using a single acoustic measure. Therefore, it is difficult to validate where such anchor stimuli lie on any unidimensional scale, including a VAS.

Finally, visual sort and rate (VSR; Granqvist, 2003) is a relatively novel rating task that retains the benefits of using a continuous scale, such as a VAS, while also encouraging the listener to make comparisons to external stimuli rather than relying on internal standards. In VSR, a listener is presented with a set of voice samples, and first sorts them into the relative order by comparing each sample one to the other. Then, the listener is asked to rate the samples along a continuum of voice quality using a VAS. By listening to multiple samples and rating them as a set, each voice serves as a comparison for all other samples in the set. The advantage over anchor methods is that no predetermined severity levels for the anchors need to be defined; they are simply judged relative to the other voices by that listener. The limited number of voices in the set and the use of a continuous rating scale (in addition to the rank ordering that is accomplished by the sorting task) provides an advantage over traditional paired comparison paradigms.

Although VSR has increasingly been used as a rating tool to generate auditory-perceptual judgments in research

settings (e.g., Gerratt et al., 2016; Heller Murray et al., 2016; Lien et al., 2015; Signorello et al., 2016), examination of the effects of VSR on reliability of these judgments relative to more frequently used rating tasks has been limited. To date, the impact of VSR on listener reliability has been explicitly examined in one study (Granqvist, 2003), with a homogenous set of voice samples (recordings from a single speaker with vocal fold nodules and a set of synthetically generated voice samples) rated by experienced clinicians. In that study, Granqvist found that VSR improved both inter- and intrarater reliability compared to ratings made using a VAS for the highly controlled sets of stimuli used in the study. Thus, it is yet unknown whether VSR is an experimental approach that may improve listener reliability for a diverse set of naturally dysphonic samples that range in severity. The purpose of this study was to compare listeners' reliability using a relatively novel experimental scaling method, called VSR, versus the more frequently used VAS method for speakers with dysphonia that ranged in OS and BR. Because the purpose of our study was to investigate a novel method that might be used by inexperienced listeners, we selected voice dimensions that could be understood and rated with at least moderate reliability to ensure valid interpretation of our results. We also wanted to investigate dimensions that would have face validity among the majority of speakers with voice disorders. OS and BR were selected as the dimensions that would meet both criteria (Zraick et al., 2011). We hypothesized that interrater reliability would be significantly better for both OS and BR when listeners used VSR compared to VAS, due to the advantages of external comparison described above. We hypothesized that intrarater reliability would not differ significantly between the two rating methods.

Method

All procedures were approved by the University of Washington Institutional Review Board.

Stimuli Selection and Preparation

Sixty speech samples (30 males, 30 females) were initially chosen by the first author from a clinical database of speakers with voice disorders to represent a range of severities and voice qualities. Speakers were chosen with a diverse range of diagnoses and who were perceived by the first author to vary across multiple voice quality dimensions (e.g., OS, BR, roughness, strain) to represent the diverse range of voice qualities encountered in a clinical setting. Speakers who were perceived to have motor speech impairment (e.g., imprecise articulation, velopharyngeal impairment) were excluded. In this experiment, stimuli were the second sentence of the Rainbow Passage (Fairbanks, 1960). These speaker stimuli were taken from a database of voice recordings obtained in prior studies in a quiet room using a headset condenser microphone (AKG C420) routed to a digital audio recorder (Tascam DAP1) and digitized at a sampling rate of 44100 Hz, as well as recordings from a commercially available voice database (Disordered Voice Database Model 4337 [Ver. 1.03], 1994). To ensure a broad distribution of stimuli, five experienced speech-language pathologists rated all of the speech samples for OS and BR using a traditional VAS. OS was defined as a comprehensive measure of how "good" or "poor" the voice is, and BR was defined as the perception of audible air escape in the voice (Eadie & Baylor, 2006). Endpoints of the VAS ranged from 0 (*normal*) to 100 (*severe*). Fifteen percent of the samples were repeated to assess intrarater reliability using Pearson correlation coefficients (for OS, mean $r = .95$, $SD = .04$; for BR, mean $r = .92$, $SD = .09$). Interrater reliability was assessed using intraclass correlation coefficients based on a single rating ($k = 1$), absolute agreement, two-way random-effects model (for OS, $ICC(2, 1) = .83$, 95% CI [.75, .89]; for BR, $ICC(2, 1) = .71$, 95% CI [.56, .81]). Experienced listeners' average ratings of OS ranged from 2 to 96 ($M = 35$, $SD = 30$), and ratings of BR ranged from 0 to 98 ($M = 23$, $SD = 26$). Average ratings from the experienced listeners were used to distribute the stimuli into sets of 10 samples with a range of severities present in each set, as outlined in the VSR procedure (Granqvist, 2003; Lien et al., 2015). The final sets included 50 (25 male, 25 female) speakers for OS and the same number for BR. Table 1 summarizes the voice diagnoses of the speakers in each set. Twenty percent of the samples were repeated to test intrarater reliability, resulting in a total of 60 samples (six sets of 10 samples) for each dimension. Repeated samples were

Table 1. Speaker diagnoses.

a) Voice disorder diagnoses included in speech samples for overall severity ratings

Diagnosis	Number
Benign vocal fold lesions	10
Nonphonotraumatic vocal hyperfunction	2
Unilateral vocal fold paresis/paralysis	12
Bilateral vocal fold paresis	4
Laryngeal cancer	3
Papilloma	2
Other (chronic laryngitis, laryngopharyngeal reflux, Reinke's edema, glottic stenosis, subglottic stenosis, unilateral vocal fold hemorrhage, multiple diagnoses)	8
Typical speaker	9
Total	50

b) Voice disorder diagnoses included in speech samples for breathiness ratings

Diagnosis	Number
Benign vocal fold lesions	10
Nonphonotraumatic vocal hyperfunction	2
Unilateral vocal fold paresis/paralysis	12
Bilateral vocal fold paresis	4
Laryngopharyngeal reflux	2
Laryngeal cancer	3
Papilloma	2
Other (bilateral vocal fold edema, chronic laryngitis, unilateral vocal fold hemorrhage, glottic stenosis, subglottic stenosis, multiple diagnoses)	8
Typical speaker	7
Total	50

distributed across sets, resulting in one to two repeats per set, and were never presented in the same set as the original.

Participants

Twenty-two inexperienced listeners were recruited to participate in the VSR and VAS rating tasks (16 women, six men; M_{age} = 28 years, range: 19–48, SD = 8.1). Eligibility criteria included age ≥ 18 years, fluent English speaker, no reported history of hearing loss, and little or no formal coursework related to voice disorders. Listeners passed a hearing screening (pure tones were presented via headphones in a quiet room at 25 dB HL at 250, 500, 1000, 2000, and 4000 Hz, using a Garson-Stadler GSI 17 portable audiometer).

Listener Procedures

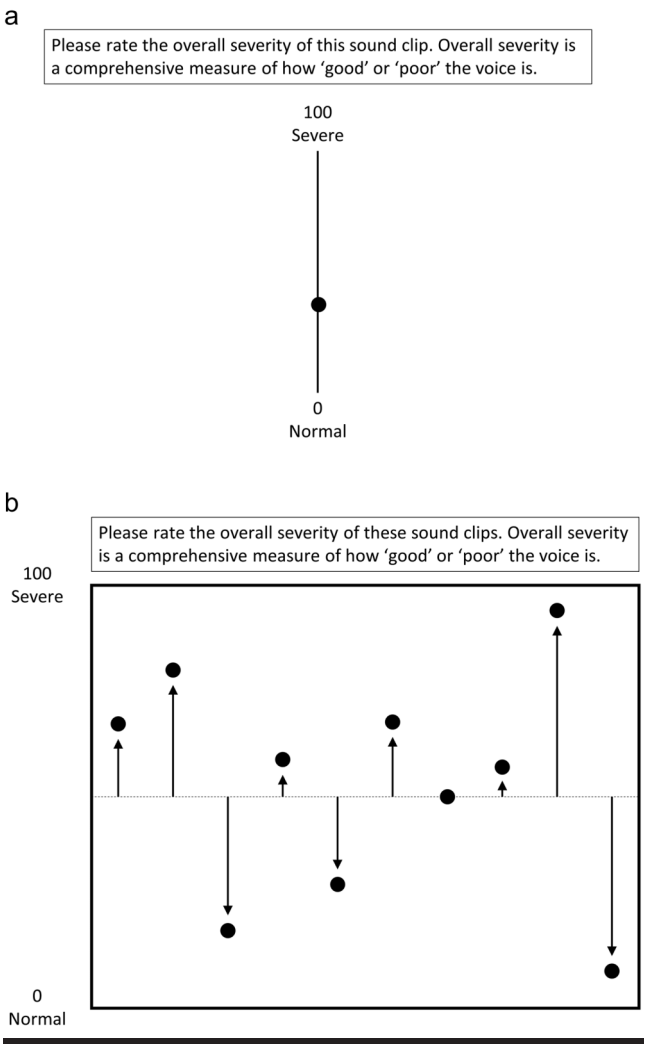
To compare the reliability of listeners' judgments of dysphonia using VSR and VAS procedures, inexperienced listeners rated the speech samples in four blocks: VSR-OS, VSR-BR, VAS-OS, and VAS-BR. The order of rating tasks (VSR/VAS) and voice quality dimensions (OS/BR) was approximately counterbalanced across participants to control for learning and order effects (22 listeners distributed across four listening order conditions). Listeners were oriented to the rating tasks with written instructions and completed a practice set comprising five voice samples (two mild, two moderate, and one severe, as selected by the first author) that were not included in the test sets before each block.

Stimuli were presented via custom-made computer programs (see Figures 1a and 1b). Both rating tasks were oriented in the vertical plane to reduce effects related to handedness (Chapanis & Gropper, 1968). For the VAS, listeners rated each sample using a vertical scroll bar with labeled endpoints (0 = *normal voice quality,* represented at the bottom of the scale; 100 = *severely dysphonic or breathy,* represented at the top of the scale; see Figure 1a). The listener moved the scroll bar to rate OS or BR. Samples were presented in a random order, and listeners rated one sample at a time. For the VSR task, listeners rated the samples in sets of 10. All listeners were presented with the same sets of voice samples; however, the order of sets was randomized within each dimension condition (OS/BR). Listeners saw a series of dots representing the samples, all initially oriented at midline. To sort the samples into their rank order, listeners moved each dot up or down the screen, which had a vertically oriented VAS on the left side (see Figure 1b). Listeners were instructed to compare each sample to one another, and then adjust the order and position accordingly. After making adjustments, the listener clicked a button to submit ratings for the entire set before moving to the next set. In both tasks, listeners were allowed to listen to each voice as many times as they wished.

Data Analysis

The effect of the task (VSR vs. VAS) on inexperienced listeners' ratings of OS and BR was measured for two dependent variables: intrarater reliability and interrater variability. Intrarater reliability was calculated using an individual listener's first and second ratings of repeated stimuli and determined using Pearson correlation coefficients. For descriptive purposes, interrater reliability was calculated using intraclass correlation coefficients, based on a single rating (k = 1), absolute-agreement, two-way random-effects model, for each of the four rating conditions. This model was chosen to estimate the reliability of a single rater compared to other raters in the sample. Interrater reliability was also measured using a variability score, as defined by Chan and Yiu (2002), hereafter referred to as interrater variability. To measure interrater variability, the difference between a listener's individual rating of a stimulus and the group mean rating of the stimulus was squared, and the squared differences were averaged across all stimuli to obtain an interrater variability score for the listener. In other words, where x is the listener's voice quality rating, v is the number of voice samples, and

Figure 1. Computer interfaces for (a) visual analog scale and (b) visual sort and rate.



n is the number of raters, interrater variability was calculated as seen in (1).

$$f(x) = \frac{\sum_{i=1}^{v} (x_v - \bar{x}_v)^2}{n}. \quad (1)$$

The interrater variability score was calculated for each listener in each rating task and rating dimension condition (i.e., VSR-OS, VAS-OS, VSR-BR, VAS-BR). The interrater variability score represents the degree of variability of the ratings made by each listener relative to other listeners; therefore, a low number indicates the listener varied less from the group average, whereas a high score indicates that a listener varied more from the group average. Interrater variability was selected as one of the primary outcome variables as it takes into account the amount of variability in individual listeners' ratings without the masking effects of group correlations, and without creating arbitrary cutoffs or imposing a linear scale upon a perceptual dimension that may behave in a nonlinear fashion, as is inherent in interrater agreement measures (Eadie & Kapsner-Smith, 2011).

The Shapiro–Wilk test of normality was conducted on the differences in the dependent variables (interrater variability scores and intrarater reliability scores) between rating conditions (VSR vs. VAS), for each perceptual dimension (OS and BR), to assess the appropriateness of use of parametric statistics. For both interrater variability and intrarater reliability scores and for both perceptual dimensions, Shapiro–Wilk tests of the differences between VAS and VSR did not show a significant departure from normality. Paired t tests were calculated to determine whether there was a significant difference related to task (VAS vs. VSR) for the two dependent variables, and effect sizes and confidence intervals were calculated. For descriptive purposes, a variability score was also calculated for each voice sample. The difference between each listener's individual rating of a voice sample and the group mean rating was squared, and these differences were averaged across all listeners, giving a variability score for each stimulus, shown in (2). Voice sample variability scores were examined for patterns of variability across severity levels of OS and BR in the two rating tasks. Statistical tests were conducted using SPSS Version 26.

$$f(x) = \frac{\sum_{i=1}^{n} (x_n - \bar{x}_n)^2}{v}. \quad (2)$$

Results

Two listeners did not complete the rating task, and as a result, their results were not included in the final analysis. Outlier data from two additional listeners were excluded because their variability scores were more than 2 *SD*s above the group mean in at least one rating task, suggesting the task was not adequately understood and results were not valid (Osborne & Overbay, 2004). Consequently, results are based on data from 18 listeners.

Intrarater Reliability

For OS, there was no significant difference between the intrarater reliability of ratings obtained using VAS (mean r = .92, SD = .08) compared to VSR (mean r = .88, SD = .12); $t(17)$ = 1.39, 95% CI [−.02, .10], p = .182, d = 0.33. In contrast, intrarrater reliability was significantly higher for BR ratings obtained using VAS (mean r = .94, SD = .04) relative to VSR (mean r = .90, SD = .08); $t(17)$ = 2.14, 95% CI [.001, .09], p = .047, d = 0.50.

Interrater Variability

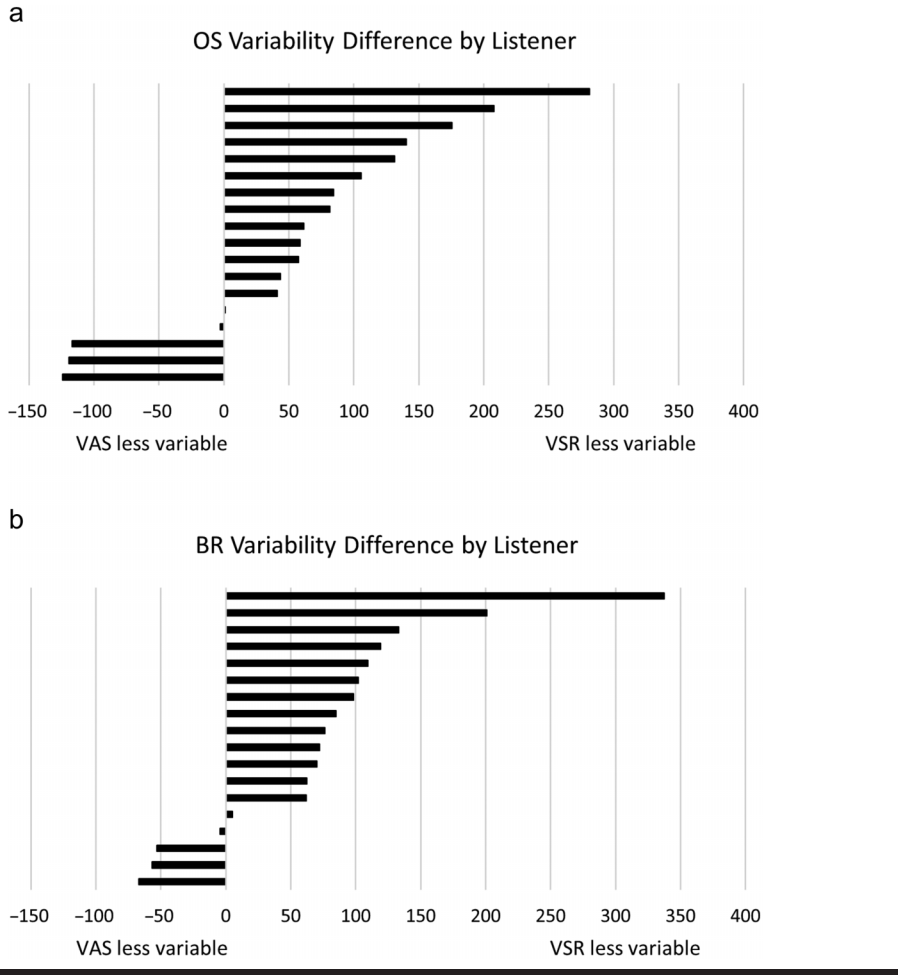
For descriptive purposes, interrater reliability was first measured using intraclass correlation coefficients, to facilitate comparisons with prior studies. The ICC estimate was calculated based on a single rating (k = 1), absolute-agreement, two-way random-effects model, for each of the four conditions: VAS-OS, $ICC(2, 1)$ = .80, 95% CI [.73, .86]; VSR-OS, $ICC(2, 1)$ = .84, 95% CI [.78, .90]; VAS-BR, $ICC(2, 1)$ = .74, 95% CI [.66, .82]; VSR-BR, $ICC(2, 1)$ = .80, 95% CI [.73, .86].

Second, interrater variability scores were calculated for statistical comparisons. These scores reflect the degree of deviation of an individual listener's ratings from the group mean ratings of each stimulus; therefore, a lower score indicates better interrater reliability. For OS, there was a significant difference between interrater variability of ratings obtained using VAS (M = 240, SD = 111) compared to VSR (M = 179, SD = 85.7); $t(17)$ = 2.39, 95% CI [7.17, 116], p = .029, d = 0.56. For BR, there was also a significant difference between interrater variability of ratings obtained using VAS (M = 300, SD = 104) compared to VSR (M = 225, SD = 72.5); $t(17)$ = 3.30, 95% CI [27.1, 123], p = .004, d = 0.78. Thus, interrater variability of ratings made using VSR were significantly lower (better) than interrater variability of ratings made using VAS for both OS and BR.

To examine whether group performance masked any individual listener effects, differences in interrater variability scores across ratings tasks (variability for VAS minus variability for VSR) for each listener are presented in Figure 2 for descriptive purposes (*OS* M = 61.6, range: −124 to 281, SD = 110; *BR* M = 75.1, range: −67.0 to 337, SD = 96.4). A positive change score indicates a reduction in variability (i.e., better interrater reliability) for VSR relative to VAS. The figure shows that, for both OS and BR, 13 (72%) listeners showed lower interrater variability for VSR than VAS; two listeners (11%) showed no clear difference, and three listeners (17%) showed lower interrater variability with VAS than VSR.

Variability scores also were calculated for each voice sample by averaging the squared differences between each listener's rating and the group average rating (Chan & Yiu, 2002; Eadie & Kapsner-Smith, 2011). Mean interrater variability scores for each speaker were plotted as a function

Figure 2. Differences in variability scores for each listener by rating task (visual analog scale [VAS] minus visual sort and rate [VSR]) for (a) overall severity and (b) breathiness. A positive value indicates a reduction in variability (i.e., better interrater reliability) for VSR compared to VAS, and a negative value indicates an increase in variability (i.e., worse interrater reliability) for VSR compared to VAS. A score of 0 indicates no difference in listener variability between the scale types. OS = overall severity; BR = breathiness.

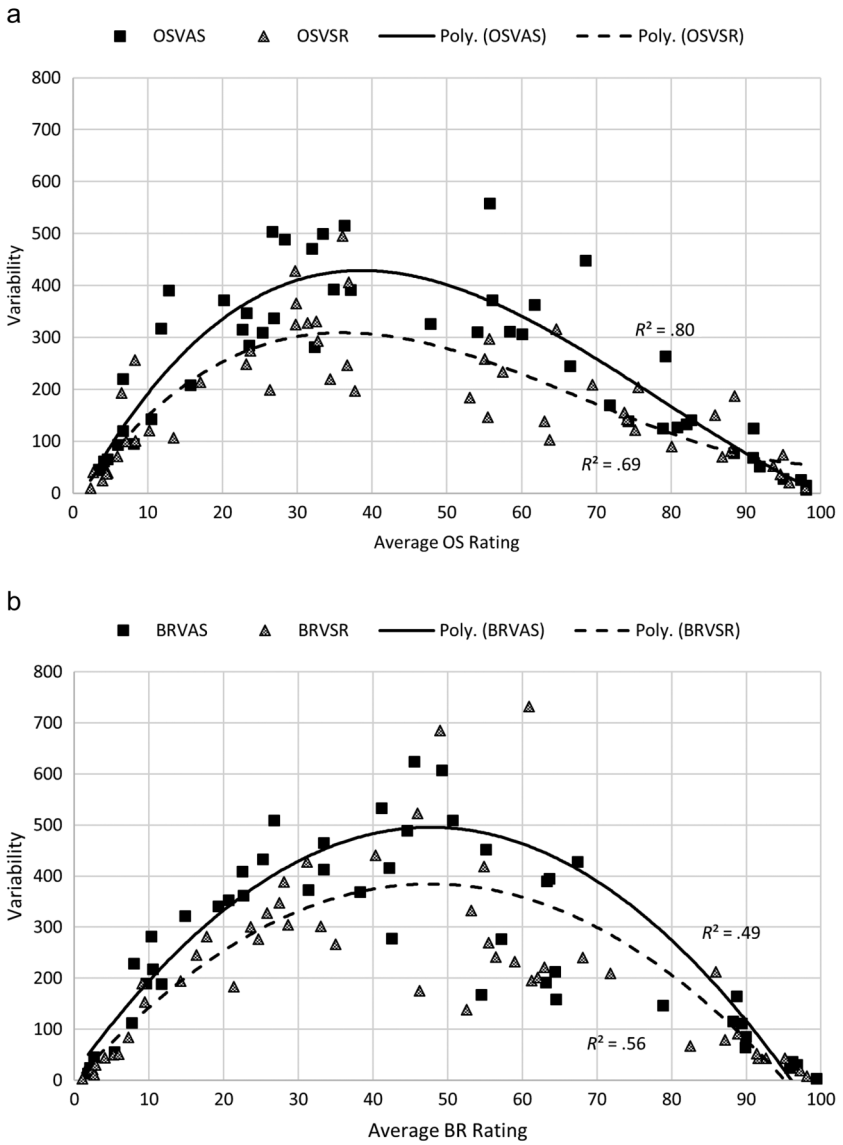


of the severity of each speaker sample for each rating task (VSR and VAS) for both OS (see Figure 3a) and BR (see Figure 3b). After visual inspection, simple regression was used to assess the relationship between severity (x) and variability (y). Sequential predictor entry was used, with each block including consecutively higher order polynomials (x, x^2, x^3, etc.) until no significant improvement in model fit was obtained. The lines of best fit for VSR and VAS are displayed in Figures 3a and 3b with corresponding R^2 values. In all cases, data were best predicted with nonlinear functions; the highest interrater variability was observed in the midrange of speaker severity, and the lowest interrater variability (i.e., better interrater reliability) was observed at the ends. For OS, a cubic model provided the best fit for VAS ($y = -32.66 + 26.83x - 0.46x^2 + 0.002x^3$) and VSR ($y = -21.98 + 20.82x - 0.39x^2 + 0.002x^3$); for BR, a quadratic model provided the best fit for VAS ($y = 13.73 + 20.20x - 0.21x^2$) and VSR ($y = -3.14 + 16.26x - 0.17x^2$). Importantly, both Figures 3a and 3b also show that, regardless of the severity of the rating dimension, ratings using VSR were less variable than ratings from VAS. Stimuli that were in the midrange of the scale showed the greatest decreases in interrater variability scores (i.e., better interrater reliability) when ratings were made using VSR, compared to VAS.

Discussion

In this study, the reliability of a relatively novel experimental method, VSR, was compared to the more frequently used method of VAS for inexperienced listeners' ratings of OS and BR of speakers with dysphonia. Overall, findings showed that the VSR method significantly reduced interrater variability for ratings of both OS and BR of dysphonic voices, with medium-to-large effect sizes. The effect was strongest for ratings of BR; this finding is unsurprising, as ratings of individual voice quality dimensions such

Figure 3. Voice sample variability scores for judgments of (a) overall severity (OS) and (b) breathiness (BR) as a function of severity and rating task, visual analog scale (VAS) versus visual sort and rate (VSR; note, for clarity of visualization, data points not pictured in panel b for VAS at (51, 1213) and (65, 1189), and for VSR at (53, 870), though all were included in regression analyses). Curves were fitted using polynomial regression for descriptive purposes.



as BR are typically more difficult for listeners to judge than comprehensive judgments of overall voice severity, as reflected by lower reliability of these judgments (Kreiman & Gerratt, 1998). Furthermore, improved interrater reliability was consistent across the majority (72%) of listeners for both dimensions. Finally, results from this study showed that the VSR method was effective in significantly reducing interrater variability across the range of scale values, particularly in the midrange of the scale where variability in listeners' judgments of voice quality tends to be highest in both research and clinical contexts (Eadie & Kapsner-Smith, 2011; Kreiman & Gerratt, 1996; Kreiman et al., 1993).

Although several studies have used VSR as a methodological approach that theoretically should strengthen the reliability of auditory-perceptual measures of voice quality (Anand & Stepp, 2015; Gerratt et al., 2016; Heller Murray et al., 2016; Kreiman et al., 2015; Lien et al., 2015; Samlan & Kreiman, 2014; Signorello et al., 2016; Zhang et al., 2013), only one study has explicitly compared the reliability of this method to VAS, using recordings of a single speaker and synthetic stimuli (Granqvist, 2003). Consistent with this study, Granqvist (2003) showed that the VSR method was significantly better than a VAS for measures of interrater reliability. This study extended the findings from Granqvist (2003) to show that inexperienced listeners were significantly less

variable in their ratings using VSR for a large number of naturally dysphonic speakers who varied across multiple voice quality dimensions. These findings suggest that VSR is an advantageous method over VAS for evaluating a wide array of voice stimuli and severities, and should be considered as an approach for measuring voice quality in research studies.

According to a model of auditory-perceptual voice quality assessment proposed by Kreiman et al. (1993), one source of variability in listener judgments is a listener's idiosyncratic internal standards for voice quality. In an unanchored VAS rating task or an *n*-point scale, listeners must rely on comparison with these internal standards to make judgments. In this study, listeners appeared to benefit from making comparisons among stimuli in the VSR task, as reflected in lower interrater variability scores, despite the diversity of voice qualities and diagnoses of the speaker samples used in this study. The advantages conferred by VSR are not unlike those for anchor samples (i.e., a set of reference stimuli with explicit positions along the scale; Awan & Lawson, 2009; Eadie & Kapsner-Smith, 2011). However, one advantage of VSR over an anchored VAS or *n*-point scale is that it is not necessary to identify naturally dysphonic anchor stimuli that differ along a single dimension at agreed upon severity levels.

VSR proved beneficial in terms of reducing interrater variability in this study; however, intrarater reliability was less clearly impacted. VAS demonstrated a significant increase in intrarater reliability of BR judgments compared to VSR, with a medium effect size ($d = 0.50$). No significant differences were observed in intrarater reliability for OS judgments between the two methods (VAS vs. VSR). If in fact listeners are relying on external comparisons rather than internal standards to make judgments in the VSR task, it is possible that repeated samples presented with different comparison voices would result in somewhat different ratings, thus reducing intrarater reliability compared to the VAS task. This effect could be more significant for individual voice quality dimensions (e.g., BR) that tend to have lower reliability (Granqvist, 2003; Kreiman & Gerratt, 1998). These results should be interpreted cautiously, however, given the overall strong intrarater reliability values for both OS and BR in this study, irrespective of the rating task (ranging from $r = .88$–94). These intrarater reliability values are relatively higher than comparable studies that have used both VAS (anchored; Awan & Lawson, 2009) and VSR (Granqvist, 2003) and may represent a ceiling effect within individual listeners in this study.

These results are promising for the use of VSR, but should be interpreted with caution in terms of generalization. This study included listeners who were unfamiliar with both rating scales and dysphonic speakers. As a result, it is unknown how the present results would generalize to experienced clinicians who make similar types of judgments using a VAS in clinical practice (Kempster et al., 2009). Likewise, although the VSR task used in this study provided a significant reduction in interrater variability compared to ratings of OS and BR using a VAS, it is also unknown whether similar advantages would be conferred for ratings of other voice quality dimensions, such as roughness or strain. However, given that ratings of OS and BR are typically among the most reliable even for inexperienced listeners (Eadie & Baylor, 2006; Patel et al., 2010; Zraick et al., 2011), we predict that the VSR method would demonstrate an even stronger effect over typical rating scales for other voice dimensions. VSR has been used with reported high reliability for inexperienced listener ratings of naturalness and intelligibility in speakers with Parkinson's disease (Anand & Stepp, 2015), perceived vocal effort in speakers simulating vocal effort (Lien et al., 2015), OS of dysphonia in synthetic stimuli (Gerratt et al., 2016), and OS and strain in those at high risk for voice disorders (Heller Murray et al., 2016), among others.

Although VSR is relatively easy to use in an experimental setting, it is not yet ready for clinical application. Future research is needed to identify the key components of the VSR task, such as the number and distribution of comparison voices required to produce a reduction in interrater variability. Psychometric properties of the VSR should also be delineated, including whether such scales produce ratio or interval level data that are necessary for validly capturing a number of voice quality dimensions (Eadie & Doyle, 2002). VSR may also be a useful tool for training inexperienced clinicians. Evidence regarding the impact of training on variability of auditory-perceptual voice quality judgments is promising (Chan & Yiu, 2002, 2006; Eadie & Baylor, 2006), and some studies suggest that provision of feedback may improve performance of inexperienced raters (Anand et al., 2019). Given that the use of VSR alone reduces interrater variability of ratings by inexperienced listeners, the combination of VSR with feedback in a rater training paradigm may also have the potential to further improve rater performance, and remains a direction for future study.

In recent years, several research groups have proposed alternatives for addressing challenges related to rater reliability for auditory-perceptual measures of voice (Eddins et al., 2020; Kreiman & Gerratt, 2005; Kreiman et al., 2007; Patel et al., 2010). Although these approaches are promising, future research is needed to establish their validity across different types of stimuli (in particular, connected speech) that also extend to speakers with different diagnoses and clinically relevant voice quality dimensions. In the meantime, we propose that VSR is another viable alternative that should be considered as an option in each researcher's methodological toolbox when perceived voice quality is the dependent variable of interest.

Conclusions

Development of protocols to reduce variability of auditory-perceptual judgments of voice quality remains an important goal for reliable measurement of voice quality. This study showed that a relatively novel method, called VSR, significantly decreased the variability of auditory-perceptual judgments between inexperienced listeners when rating speakers with a range of dysphonic severities and

disorders, when compared to the more typically used VAS. Intrarater reliability was strong overall, regardless of the rating task. Future research should determine whether a clinically viable tool may be developed based on VSR principles, and whether such benefits extend to experienced listeners.

Acknowledgments

This work was supported by Grant R01DC015570 (awarded to C. E. S.) from the National Institute on Deafness and Other Communication Disorders. Portions of this paper were presented at the Annual Fall Voice Conference, October 2018, Seattle, WA, and at the Annual Convention of the American Speech-Language-Hearing Association, November 2018, Boston, MA. Thank you to the members of the Vocal Function Lab for assistance with perceptual data analysis. Thank you to Jake Herrmann and Andrew Smith for creating the auditory-perceptual interfaces used for data collection.

References

Anand, S., Kopf, L. M., Shrivastav, R., & Eddins, D. A. (2019). Objective indices of perceived vocal strain. *Journal of Voice, 33*(6), 838–845. https://doi.org/10.1016/j.jvoice.2018.06.005

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 58*(4), 1134–1144. https://doi.org/10.1044/2015_JSLHR-S-14-0243

Awan, S. N., & Lawson, L. L. (2009). The effect of anchor modality on the reliability of vocal severity ratings. *Journal of Voice, 23*(3), 341–352. https://doi.org/10.1016/j.jvoice.2007.10.006

Chan, K. M., & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45*(1), 111–126. https://doi.org/10.1044/1092-4388(2002/009)

Chan, K. M., & Yiu, E. M. (2006). A comparison of two perceptual voice evaluation training programs for naive listeners. *Journal of Voice, 20*(2), 229–241. https://doi.org/10.1016/j.jvoice.2005.03.007

Chapanis, A., & Gropper, B. A. (1968). The effect of the operator's handedness on some directional stereotypes in control-display relationships. *Human Factors, 10*(4), 303–320. https://doi.org/10.1177/001872086801000401

Disordered Voice Database Model 4337 (Ver. 1.03). (1994). Kay Elemetrics.

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20*(4), 527–544. https://doi.org/10.1016/j.jvoice.2005.08.007

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *The Journal of the Acoustical Society of America, 112*(6), 3014–3021. https://doi.org/10.1121/1.1518983

Eadie, T. L., & Doyle, P. C. (2005). Classification of dysphonic voice: Acoustic and auditory-perceptual measures. *Journal of Voice, 19*(1), 1–14. https://doi.org/10.1016/j.jvoice.2004.02.002

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)

Eddins, D. A., Anand, S., Lang, A., & Shrivastav, R. (2020). Developing clinically relevant scales of breathy and rough voice quality. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2019.12.021

Fairbanks, G. (1960). *Voice and articulation drillbook* (Vol. 127). Harper.

Gerratt, B., Kreiman, J., & Garellek, M. (2016). Comparing measures of voice quality from sustained phonation and continuous speech. *Journal of Speech, Language, and Hearing Research, 59*(5), 994–1001. https://doi.org/10.1044/2016_JSLHR-S-15-0307

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology, 28*(3), 109–116. https://doi.org/10.1080/14015430310015255

Heller Murray, E. S., Hands, G. L., Calabrese, C. R., & Stepp, C. E. (2016). Effects of adventitious acute vocal trauma: Relative fundamental frequency and listener perception. *Journal of Voice, 30*(2), 177–185. https://doi.org/10.1016/j.jvoice.2015.04.005

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21*(5), 576–590. https://doi.org/10.1016/j.jvoice.2006.05.001

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

Kreiman, J., Garellek, M., Chen, G., Alwan, A., & Gerratt, B. R. (2015). Perceptual evaluation of voice source models. *The Journal of the Acoustical Society of America, 138*(1), 1–10. https://doi.org/10.1121/1.4922174

Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *The Journal of the Acoustical Society of America, 100*(3), 1787–1795. https://doi.org/10.1121/1.416074

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104*(3, Pt. 1), 1598–1608. https://doi.org/10.1121/1.424372

Kreiman, J., & Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America, 117*(4 Pt. 1), 2201–2211. https://doi.org/10.1121/1.1858351

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122*(4), 2354–2364. https://doi.org/10.1121/1.2770547

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

Latoszek, B. B. V., Maryn, Y., Gerrits, E., & De Bodt, M. (2018). A meta-analysis: Acoustic measurement of roughness and breathiness. *Journal of Speech, Language, and Hearing Research, 61*(2), 298–323. https://doi.org/10.1044/2017_JSLHR-S-16-0188

Lien, Y. S., Michener, C. M., Eadie, T. L., & Stepp, C. E. (2015). Individual monitoring of vocal effort with relative fundamental frequency: Relationships with aerodynamics and listener perception. *Journal of Speech, Language, and Hearing Research, 58*(3), 566–575. https://doi.org/10.1044/2015_JSLHR-S-14-0194

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice, 24*(5), 540–555. https://doi.org/10.1016/j.jvoice.2008.12.014

Myles, P. S., Troedel, S., Boquest, M., & Reeves, M. (1999). The pain visual analog scale: Is it linear or nonlinear? *Anesthesia & Analgesia, 87*(6), 1517–1520. https://doi.org/10.1213/00000539-199912000-00038

Nagle, K. F. (2016). Emerging scientist: Challenges to CAPE-V as a standard. *Perspectives of the ASHA Special Interest Groups, 1*(3), 47–53. https://doi.org/10.1044/persp1.SIG3.47

Nemr, K., Simões-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). GRBAS and CAPE-V scales: High reliability and consensus when applied at different times. *Journal of Voice, 26*(6), 812.e17–822.e22. https://doi.org/10.1016/j.jvoice.2012.03.005

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation, 9*(1), Article 6.

Patel, S., Shrivastav, R., & Eddins, D. A. (2010). Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice, 24*(2), 168–177. https://doi.org/10.1016/j.jvoice.2008.08.002

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain, 17*(1), 45–56. https://doi.org/10.1016/0304-3959(83)90126-4

Samlan, R. A., & Kreiman, J. (2014). Perceptual consequences of changes in epilaryngeal area and shape. *The Journal of the Acoustical Society of America, 136*(5), 2798–2806. https://doi.org/10.1121/1.4896459

Signorello, R., Zhang, Z., Gerratt, B., & Kreiman, J. (2016). Impact of vocal tract resonance on the perception of voice quality changes caused by varying vocal fold stiffness. *Acta Acustica united with Acustica, 102*(2), 209–213. https://doi.org/10.3813/AAA.918937

Stevens, S. (1975). *Psychophysics*. Wiley.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review, 101*(2), 266. https://doi.org/10.1037/0033-295X.101.2.266

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice, 13*(4), 508–517. https://doi.org/10.1016/S0892-1997(99)80006-X

Zealley, A., & Aitken, R. (1969). *A growing edge of measurement of Feelings [Abridged]: Measurement of mood*. SAGE. https://doi.org/10.1177/003591576906201006

Zhang, Z., Kreiman, J., Gerratt, B. R., & Garellek, M. (2013). Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. *The Journal of the Acoustical Society of America, 133*(1), 453–462. https://doi.org/10.1121/1.4770235

Zraick, R., 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)