JSLHR

Research Article

# Individual Monitoring of Vocal Effort With Relative Fundamental Frequency: Relationships With Aerodynamics and Listener Perception

Yu-An S. Lien,[a] Carolyn M. Michener,[a] Tanya L. Eadie,[b] and Cara E. Stepp[a]

**Purpose:** The acoustic measure relative fundamental frequency (RFF) was investigated as a potential objective measure to track variations in vocal effort within and across individuals.

**Method:** Twelve speakers with healthy voices created purposeful modulations in their vocal effort during speech tasks. RFF and an aerodynamic measure of vocal effort, the ratio of sound pressure level to subglottal pressure level, were estimated from the aerodynamic and acoustic signals. Twelve listeners also judged the speech samples for vocal effort using the visual sort and rate method.

**Results:** Relationships between RFF and both the aerodynamic and perceptual measures of vocal effort were weak across speakers ($R^2$ = .06–.26). Within speakers, relationships were variable but much stronger on average ($R^2$ = .45–.56).

**Conclusions:** RFF showed stronger relationships between both the aerodynamic and perceptual measures of vocal effort when examined within individuals versus across individuals. Future work is necessary to establish these relationships in individuals with voice disorders across the therapeutic process.

*Vocal hyperfunction* is defined as “conditions of abuse and/or misuse of the vocal mechanism due to excessive and/or ‘imbalanced’ muscular forces” (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1989, p. 373), resulting in a voice that is often described as highly effortful and/or excessively strained. For such conditions, one-time assessments in the clinic may not be sufficient to accurately characterize the behavior, and long-term voice monitoring can be used to provide further insight (Hillman & Mehta, 2011). In addition to being used for initial assessment, voice monitoring also enables clinicians to track changes in a client’s voice over time, allowing for evaluation of both client adherence to behavioral voice changes as well as the effectiveness of ongoing therapy.

Voice monitoring can be based exclusively on clients’ self-reports, including self-rated voice quality. However, clients’ self-reports of voice quality have been found to show poor agreement with clinicians’ and inexperienced listeners’ evaluations (Eadie et al., 2010; Lee, Drinnan, & Carding, 2005). The perception of vocal effort or strain is particularly problematic: The test–retest reliability for clients’ self-agreement on their perceived strain is moderate, but when compared with clinician evaluation, the level of agreement is worse than chance (Lee et al., 2005). However, even within clinicians, evaluation of strain may be unreliable, depending upon the methods used and the experience of the listeners (De Bodt, Wuyts, Van de Heyning, & Croux, 1997; Eadie et al., 2010; Granqvist, 2003; Wuyts, De Bodt, & Van de Heyning, 1999).

The importance of vocal effort as an outcome measure and the variability of auditory-perceptual methods have led some researchers to seek out methods that may be more reliable for measuring this construct, such as acoustic measures (Rosenthal, Lowell, & Colton, 2014). Unfortunately, a study examining the correlation between 19 common acoustic measures and a clinician-based perceptual voice assessment protocol, the Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) scale (Hirano, 1981), found that no acoustic measure was well correlated with strain (Bhuta, Patrick, & Garnett, 2004). More recently, the perception of strain has been shown to have a strong relationship with

[a]Boston University, MA
[b]University of Washington, Seattle

Correspondence to Yu-An S. Lien: slien@bu.edu

Editor: Jody Kreiman
Associate Editor: Scott Thomson

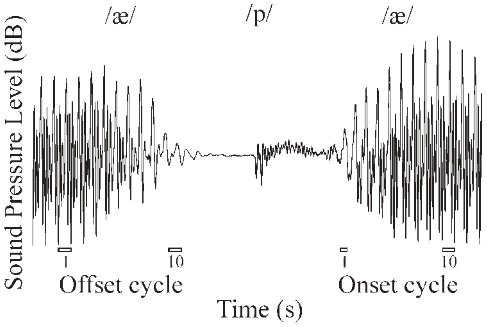
Received July 13, 2014
Revision received October 6, 2014
Accepted January 9, 2015
DOI: 10.1044/2015_JSLHR-S-14-0194

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

cepstral measures and moderate relationship with spectral measures in a group of dysphonic speakers with predominantly strained voice quality (Lowell, Kelley, Awan, Colton, & Chan, 2012). However, when the primary factor of the cepstral measure—cepstral peak prominence—was examined in a group of individuals with nonhomogeneous diagnoses, no significant correlation was found between cepstral peak prominence and the perception of strain (Brinca, Batista, Tavares, Gonçalves, & Moreno, 2014).

Cepstral measures have been shown to highly correlate with dysphonia severity (Awan & Roy, 2006); thus, a possible explanation for the disparity between the study by Lowell et al. (2012) and Brinca et al. (2014) may be that in the former study, cepstral measures were correlating with overall severity rather than with strain. Alternatively, studies also suggest that a new acoustic measure, relative fundamental frequency (RFF), may be adapted for the assessment of strain or vocal effort. RFF is measured from a voiced–voiceless consonant–voiced speech sequence (see Figure 1) and is defined as the 10 normalized fundamental frequencies immediately preceding and following the voiceless consonant, measured in semitones (ST). The RFF estimated from the offset of the vowel preceding the voiceless consonant is referred to as the *offset RFF*, and the RFF estimated from the onset of the vowel following the voiceless consonant is referred to as the *onset RFF*. RFF has been shown to differ between individuals with and without vocal hyperfunction (Stepp, Hillman, & Heaton, 2010). In young individuals with healthy voices, offset RFF tends to remain around 0 ST or decrease slightly as a function of cycle, reaching a final value of −0.84 to 0.44 ST, whereas onset RFF tends to decrease sharply as a function of cycle, starting at an initial value of 2.3 to 2.8 ST (Robb & Smith, 2002; Watson, 1998). In contrast, in individuals with vocal hyperfunction, both offset and onset RFF tend to be lower in comparison with individuals with healthy voices. The RFF of individuals with vocal hyperfunction normalizes after successful voice therapy (Stepp, Merchant, Heaton, & Hillman, 2011) but has been found not to significantly change following surgery (Stepp et al., 2010). These results suggest that the measure is sensitive to the functional nature of vocal hyperfunction, which may or may not be accompanied by organic pathologies (Hillman et al., 1989).

**Figure 1.** An acoustic waveform of a voiced-voiceless consonant-voiced instance, /æpæ/, is shown. Relative fundamental frequency (RFF) in semitones (ST) can be estimated by normalizing the 10 instantaneous fundamental frequencies in the offset vowel preceding the voiceless consonant and the onset vowel following the voiceless consonant by a steady-state fundamental frequency. The steady-state fundamental frequencies for the offset cycles and onset cycles are taken from the first offset cycle and the 10th onset cycle, respectively. The bars denote the first and 10th cycles of the offset and onset vowels.



Although RFF has shown promise for the assessment of vocal hyperfunction, previous studies have examined this measure only across participants. Furthermore, the effectiveness of RFF compared with other, relatively more established measures of vocal effort is unknown. Here, we compared RFF with two measures within participants who are creating purposeful modulations in their vocal effort to determine the usefulness of RFF for tracking variations in vocal effort in individuals. We also contrasted the usefulness of tracking vocal effort with RFF within individuals to that of tracking across individuals. Previous work by Rosenthal et al. (2014) has shown that individuals can create these types of purposeful modulations in their vocal effort and that these fluctuations correspond to changes in acoustic and aerodynamic parameters of their voice as well as a listener's perception of vocal quality. However, relationships among these measures were not examined. Thus, here we compared the acoustic measure RFF with both an aerodynamic measure involving subglottal (tracheal) pressure as well as a measure of listener perception of vocal effort.

The typical aerodynamic profile observed in individuals with vocal hyperfunction relative to individuals with typical voices is increased subglottal pressure (Hillman et al., 1989; Hillman, Montgomery, & Zeitels, 1997; Netsell, Lotz, & Shaughnessy, 1984). It has been hypothesized that individuals with vocal hyperfunction require higher subglottal pressure (driving pressure) to achieve phonation due to their heightened levels of muscle tension, increased vocal fold stiffness, underadduction, and/or hyperadducted laryngeal airway (Hillman et al., 1989; Netsell et al., 1984). However, subglottal pressure can be used for the assessment of vocal hyperfunction (Hillman et al., 1997) only when sound pressure level is accounted for, given that increases in subglottal pressure are highly correlated with increases in sound pressure level and are often used as a strategy to increase loudness (Ladefoged & McKinney, 1963).

Although subglottal pressure is appropriate for clinical settings, it is unsuitable for remote individual tracking of vocal effort due to the need for specialized equipment and the training required to accurately collect and estimate the measure noninvasively. In addition, the aerodynamic profile may be confounded by the presence of vocal fold lesions (e.g., nodules, polyps, contact ulcers) because the aerodynamic profile may vary depending on the type of vocal fold lesion, the structure and location of the lesion, and the particular compensatory mechanism developed (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1990). Here, we compared the relationship between RFF and the aerodynamic ratio of sound pressure level with subglottal pressure level within and across individuals. This ratio was chosen because it can detect increases in subglottal pressure due to increased hyperfunctional voice quality yet normalizes for potential changes in sound pressure level.

In addition to comparing RFF with aerodynamic measures of vocal effort, we also compared RFF with listeners' perceptions of vocal effort. RFF was found to show significant yet weak correlations with listener perception of vocal effort in a previous study that compared the two measures across individuals (Stepp, Sawin, & Eadie, 2012). However, rating across multiple speakers who may differ in multiple dimensions of vocal quality could confound listeners' judgments on the single dimension of vocal effort. In addition, individual speakers may vary in their typical voice quality. Thus, we hypothesize that we will observe a stronger relationship between RFF and listener perception of vocal effort when it is examined within individuals compared with across individuals. Comparison of RFF with both an aerodynamic measure and a perceptual measure of vocal effort within speakers will aid in determining the effectiveness and reliability of RFF for individual tracking of vocal effort.

## Method

### *Experiment 1: Relationship Between RFF and an Aerodynamic Measure of Vocal Effort*

*Speakers.* Twelve young adult speakers (seven women and five men, $M$ = 22 years, $SD$ = 2.7 years) participated in this study. All participants were native speakers of American English, nonsmokers, and had no prior history of speech, language, or hearing disorders. One of the participants had previously received professional singing training. The participants completed written consent in compliance with the Boston University Institutional Review Board.

*Experimental protocol.* All recordings took place in a sound-treated room. Each speaker was instructed to produce two /pæ/ trains, each consisting of seven /pæ/ productions, first using his or her own typical loudness and pitch. The participants were trained to produce each /pæ/ train in a connected, legato fashion in a single exhalation (Plexico, Sandage, & Faver, 2011). The intraoral air pressure, oral airflow, and sound pressure levels were recorded using the Phonatory Aerodynamic System (Model 6600, KayPENTAX, Lincoln Park, NJ). The intraoral air pressure during the /p/ occlusion was used to approximate the subglottal pressure (Hertegård, Gauffin, & Lindestad, 1995). An experimenter monitored the participant during the task and asked the participant to repeat any /pæ/ trains that were misarticulated, obviously glottalized, or had unstable measurements.

Speakers were subsequently asked to produce the same stimuli (two /pæ/ trains) using four additional levels of vocal effort relative to their typical productions in the following order: relaxed, slightly strained, moderately strained, and maximally strained. To ensure that the speakers were modulating their vocal effort, the participants were given feedback from the experimenter and real-time visual feedback based on their intraoral pressure and sound pressure level. Speakers were instructed to decrease their intraoral air pressure while keeping the sound pressure level constant to relax their voice and to increase their intraoral pressure while keeping the sound pressure level constant to increase the strain in their voice. The baseline intraoral pressure was slightly different for each speaker. In general, speakers were encouraged to aim for an intraoral pressure difference of 5 cm $H_2O$ between each vocal effort level. For example, if a speaker's typical intraoral pressure was 7 cm $H_2O$, the speaker was asked to target 12 cm $H_2O$ when producing the slightly strained voice. However, the targets for relaxed voice could not be set in this way, so speakers were asked to decrease their intraoral pressure as much as possible to produce the relaxed voice. The task was repeated until both the experimenter and participant were satisfied with the productions.

*Data analysis.* To determine the aerodynamic ratio of sound pressure level to average intraoral pressure (dB SPL/cm $H_2O$), a single investigator (C.M.M.) used KayPENTAX Phonatory Aerodynamic System Software (PAS 6600, Version 3.4) to estimate the intraoral pressure, average oral airflow, and sound pressure level during productions. The productions were rejected if either the magnitude of the oral airflow was nonzero or if the intraoral pressure peak was not flat during the production of the /p/. Adequate estimates of intraoral pressure can be obtained by averaging the middle five /pæ/ productions (Faver, Plexico, & Sandage, 2012); thus, intraoral pressure and sound pressure level were estimated using these productions from both /pæ/ trains. The sound pressure level and intraoral pressure from these nominal 10 productions ($M$ = 9.7, $SD$ = 1.1) were averaged to determine the mean dB SPL/cm $H_2O$ for each speaker at each level of vocal effort.

The investigator (C.M.M) used Praat (Version 5.3.04; Boersma & Weenink, 2012) acoustic analysis software and Microsoft Excel (Version 14) to perform the RFF analysis. First, the investigator visualized the acoustic waveform using Praat and verified that both sonorants surrounding the voiceless consonants were not glottalized. Glottalized samples were excluded due to their irregular vibratory patterns. If the sample was usable, the investigator proceeded to determine the instantaneous fundamental frequencies (F0), the inverse of the periods, of 10 vocal cycles preceding and following the voiceless consonant. An increase in subglottal pressure is known to be associated with an increase in F0 (Titze, 1989). Thus, in order to reduce the effect of increases in F0 due to increases in subglottal pressure and individual differences in baseline pitch and intrinsic pitch of vowels, these instantaneous F0 were normalized to the reference fundamental frequencies ($F0_{ref}$) in ST using Equation 1. The $F0_{ref}$ used in the calculation of the offset RFF was the F0 for the first offset cycle, and the one used in the calculation of the onset RFF was the F0 for the 10th (last) onset cycle. These $F0_{ref}$ were selected because they are the cycles farthest from the voiceless consonant and closest to the mid-portion of the vowel. Thus, these cycles are most likely to capture the changes in instantaneous F0 during devoicing and revoicing and to be at steady state. In addition, to ensure that the sonorants were at steady state, the investigator rejected the offset or onset RFF if the RFF magnitude for the second offset or ninth onset cycles (the cycles next to the reference cycle) was greater than 0.8 ST.

$$\mathrm{RFF}\ (\mathrm{ST}) = 39.86 \times \log_{10}\left(\mathrm{F0} / \mathrm{F0}_{ref}\right) \qquad (1)$$

Each sequence contained six /æpæ/ productions appropriate for RFF estimation, but in this study, the last production was excluded from the analysis because glottalization tends to occur at the end of the sequences. Consequently, RFF was estimated from nominally 10 /æpæ/ productions (offset: $M$ = 9.0, $SD$ = 2.1; onset: $M$ = 9.2, $SD$ = 1.8) in the two /pæ/ trains and averaged to calculate the RFF mean for each speaker at each effort level.

Reliability was calculated using both Pearson product–moment correlations and mean square errors. Because correlations do not indicate the degree to which any two measures agree or vary on an absolute scale, mean square errors also were included. To determine the interrater reliability, a second investigator (Y.S.L.) independently reanalyzed more than 15% of the RFF samples. The Pearson product–moment coefficient and mean square error ($MSE$) were calculated, yielding $r$ = .91 and $MSE$ = 0.22 ST. In addition, the initial investigator (C.M.M.) reanalyzed more than 15% of the samples three months after the original analysis to determine the intrarater reliability. The Pearson product–moment correlation coefficient and mean square error were calculated, yielding $r$ = .97 and $MSE$ = 0.09 ST.

All statistical analyses were completed with Minitab statistical software (Version 16.2.2; Minitab Inc., State College, PA). RFF patterns for each vocal effort level were visualized to determine whether speakers modulated their RFF while altering the vocal effort level. The relationships between the aerodynamic ratio and RFF across speakers and within speaker were examined using a coefficient of determination ($R^2$) between the aerodynamic ratio and RFF to determine the amount of variance in the aerodynamic ratio explained by RFF. Only the cycles closest to the voiceless consonant (Offset Cycle 10 RFF and Onset Cycle 1 RFF) were used in the analysis because these cycles exhibited the greatest difference between individuals with vocal hyperfunction and individuals with healthy voices in previous work (Stepp et al., 2010).

## *Experiment 2: Relationship Between RFF and a Perceptual Measure of Vocal Effort*

*Speakers.* The same speakers from Experiment 1 participated in Experiment 2 as part of the same visit. Each speaker was instructed to read sentence stimuli in the same five levels of vocal effort from Experiment 1 in the following order: typical, relaxed, slightly strained, moderately strained, and maximally strained. The stimuli were the sentences "The new pony loved wee Penny and lovely Polly as well" and "Lovely Pamela is your pal when you play more." Both sentences contained three RFF instances with the phoneme /p/ and were designed to place the /p/ between stressed vowels (Lien, Gattuccio, & Stepp, 2014). The experimenter monitored the subject for obvious misarticulations or glottalizations. When those occurred, the participant was asked to repeat the sentence. The sentences were recorded using a head-mounted microphone (Model PC131; Sennheiser, Old Lyme, CT) connected to a digital audio recorder (Model LS-10; Olympus, Center Valley, PA), and the sampling rate and resolution were 44.1 kHz and 16 bit, respectively.

*Experimental setup.* The recordings of the two sentences recorded from the 12 speakers at five different levels of vocal effort yielded a total number of 120 recordings for evaluation. Each sample was normalized for peak intensity using MATLAB (Version R2012a; MathWorks, Natick, MA).

Although inexperienced listeners' perceptions of vocal effort tend to be unreliable, studies have shown that the use of anchors and certain ratings methods can improve reliability by counteracting the effect of internal standards and facilitating comparisons across voice samples (Chan & Yiu, 2002; Granqvist, 2003). Anchors are difficult to implement because it is impossible to determine where the anchor should be positioned on the rating scale (Granqvist, 2003). Thus, in this study, we introduced familiarization samples and implemented the visual sort and rate method (Granqvist, 2003) to optimize reliability.

Three individuals with experience listening to disordered voices (pilot listeners) individually listened to all the speech samples and sorted them into five categories (L easy, L typical, L slightly strained, L moderately strained, and L maximally strained) based on the perceived level of vocal effort. The L denotes that these vocal effort levels are based on the pilot listeners' judgments. After sorting the speech samples, each pilot listener ensured that the samples had been sorted into the correct category by listening to the samples in the order sorted. That is, each pilot listener first listened to all samples in the easy category, followed by all the samples in the typical category, and so on. In this way, each stimulus becomes an external reference for the remaining stimuli, resulting in a task that is more similar to paired comparisons (or anchors) than when typical visual analog scales are used. All pilot listeners were informed that they did not need to have the same number of samples in each category.

The categories were converted into a pilot score from 1 to 5 (1 = *L easy*, 2 = *L typical*, 3 = *L slightly strained*, 4 = *L moderately strained*, and 5 = *L maximally strained*). For each speech sample, an average score across all pilot listeners was calculated. Based on the rounded average pilot scores (1–5), all speech samples were given an average pilot vocal effort rating (L easy–L maximally strained).

These pilot listening scores were used to design the rank and sort perceptual experiment (Granqvist, 2003) for this study. In the perceptual experiment, 33% of the samples (i.e., 40 samples out of 120 samples) were repeated for evaluation of intrarater reliability. Because there were five vocal effort levels (L easy–L maximally strained), eight samples were randomly selected within each category to be repeated for the measure of intrarater reliability. A total of 160 samples (120 original + 40 repeated) were divided into 20 sets, each with eight speech samples to be rated.

Within each set, the distribution of pilot vocal effort ratings of the speech samples was arranged to be similar to the overall distribution of the stimuli. The overall distribution (including the samples used for intrarater reliability) of L easy: L typical: L slightly strained: L moderately strained: L maximally strained was 21:48:42:28:21; thus, distribution of the stimuli in each set was approximately 1.1:2.4:2.1:1.4:1.1. The ratios were not integers, so the number of recordings from each category was not fixed by set. For example, most sets had one recording in the easy category, but a few had two. The speech samples from each vocal effort level were pseudorandomly assigned into sets, but no stimuli in the same set were spoken by the same person at the same attempted vocal effort level. The order of the sets and the samples in each set were randomized for each listener.

*Listeners.* Twelve inexperienced young adult listeners (six women and six men, $M$ = 22.0 years, $SD$ = 2.7 years) participated in a single visit in which they rated 20 sets of stimuli. Participants in this group were native speakers of American English and had reported no prior history of speech, language, or hearing disorders. Listeners had no prior experience with or coursework in voice disorders, formal exposure to individuals with voice disorders, or experience using rating scales for judging dysphonia. The listeners completed written consent in compliance with the Boston University Institutional Review Board.

*Experimental protocol.* First, each participant was asked to listen to the familiarization samples, which comprised six female and six male voice samples from a different dataset. Each sample contained one of two sentences ("The new pony loved wee Penny and lovely Polly as well" or "Lovely Pamela is your pal when you play more") at one of three levels of vocal effort (relaxed, slightly strained, and maximally strained). These samples were used to allow the inexperienced listeners to familiarize themselves with the different levels of vocal effort that they might hear in the study.

Listeners were then asked to use the visual sort and rate method (Granqvist, 2003) to score the sound clips from 0 to 100 based on the perceived vocal effort. *Vocal effort* was defined as "the perceived effort during phonation" (Verdolini, Titze, & Fennell, 1994, p. 1001), in which scores of 0 and 100 represented the least and most effortful voices imagined, respectively. Speech samples were presented and played over headphones adjusted to a comfortable listening level. A custom-designed computer software program was developed to present the speech samples and obtain the perceptual ratings. The visual sort and rate method (Granqvist, 2003) requires each participant to listen to each speech sample within a set and then sort the stimuli by moving them up and down on a computer screen so that icons of the speech samples are rank ordered from least (bottom of the screen) to most (top of the screen) effortful. Samples with similarly rated vocal effort lie close to each other. This method facilitates comparison between stimuli and provides an external reference during the task, which improves listener reliability for the task (Granqvist, 2003). The vertical axis of the screen consists of a 100-mm visual analog scale, such that the ranked locations of the stimuli are then fine-tuned by listeners and correspond to a 0 to 100 rating (0 = *least effortful* and 100 = *most effortful*). After 10 sets, the listeners were asked to take a mandatory 10-minute break to reduce fatigue effects. The average duration of the entire experiment was less than 1.5 hours.

*Data analysis.* The perceptual data were averaged across the two sentences and 12 listeners to generate a mean score for each speaker at each vocal effort level. Reliability for these measures was reported using Pearson product–moment correlation coefficients as well as the average standard deviation of the listeners' judgments. The averaged intrarater reliability calculated using the Pearson product–moment correlation coefficient yielded $r$ = .93; the average standard deviation within listeners yielded $SD$ = 5.54. Interrater reliability was assessed using the intraclass correlation coefficient, type 2k (Shrout & Fleiss, 1979) yielding ρ = .97; the average standard deviation among listeners yielded $SD$ = 13.1.

Using the acoustic recordings, a single investigator (C.M.M.) estimated the sentence-level RFF mean and standard deviation from the three RFF instances in a sentence using the acoustic analysis procedure described previously. The interrater reliability evaluated using the Pearson product–moment correlation coefficient and mean square error were $r$ = .88 and $MSE$ = 0.31 ST. The intrarater reliability computed using the Pearson product–moment correlation coefficient and mean square error were $r$ = .97 and $MSE$ = 0.07 ST.

Similar to previous analyses, the RFF means estimated during the /p/ sentences were visualized for each vocal effort level to determine whether speakers modulated their RFF while altering the vocal effort level. The relationships between the perceptual ratings and RFF across speakers and within speakers were examined using the coefficient of determination ($R^2$) between the perceptual ratings and RFF to determine the amount of variance in the perceptual ratings explained by RFF. Again, only the cycles closest to the voiceless consonant (Offset Cycle 10 RFF and Onset Cycle 1 RFF) were used in the analysis.

## Results

*Experiment 1.* RFF means during the /pæ/ productions are plotted as a function of cycle for each speaker vocal effort level in the top panel of Figure 2. During the relaxed, typical, and slightly strained conditions, offset

**Figure 2.** RFF means in semitones (ST) estimated from the /pæ/ productions (top) and from the /p/ sentences (bottom) are plotted as a function of cycle (Offset Cycles 1–10 and Onset Cycles 1–10) for each vocal effort level (relaxed, typical, strain1—slightly strained, strain2—moderately strained, strain3—maximally strained). Error bars indicate the 95% confidence intervals for the means.

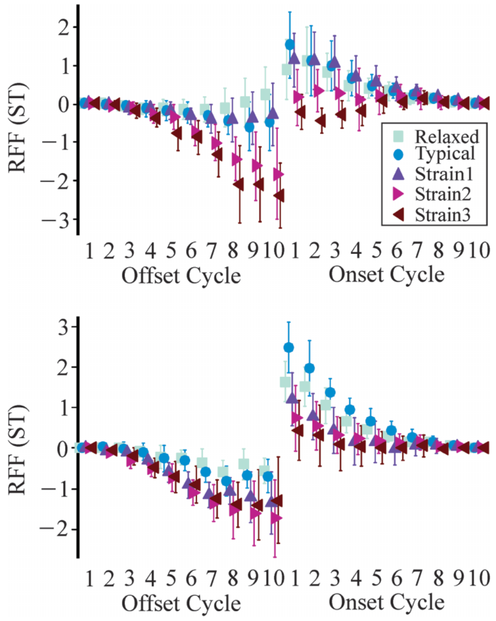
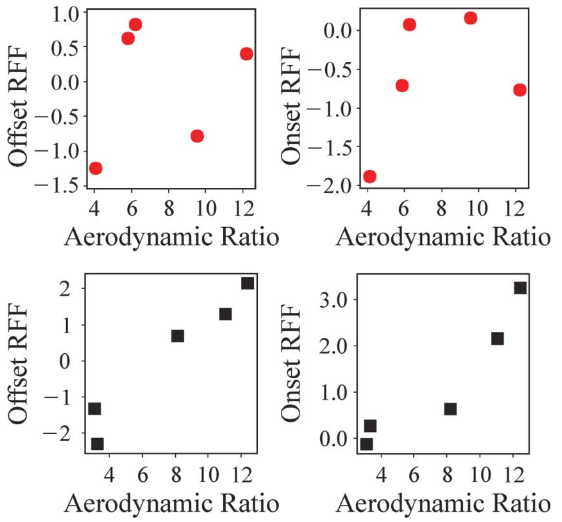


**Figure 3.** Offset Cycle 10 (left) and Onset Cycle 1 (right) RFF as a function of the aerodynamic ratio. Top: a participant whose productions yielded low $R^2$ between RFF and the aerodynamic ratio. Bottom: a participant whose productions yielded high $R^2$ between RFF and the aerodynamic ratio.

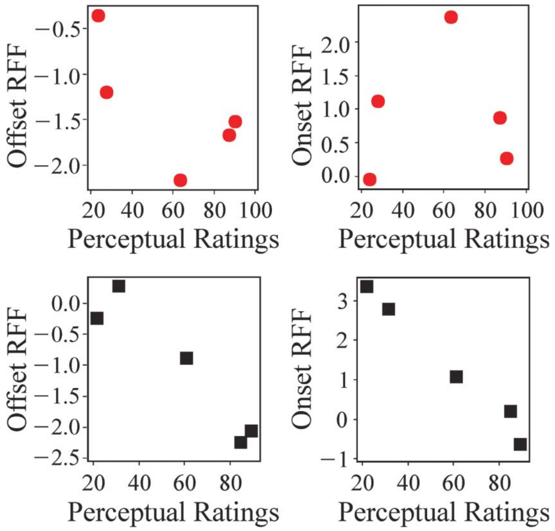


RFF remained around 0 ST for all cycles and onset RFF decreased sharply as a function of cycle. The RFF for the moderately strained and the maximally strained conditions was generally lower: The offset RFF decreased as a function of cycle and the onset RFF was steady or even increased slightly as a function of cycle.

When examined across speakers, the relationship between RFF and the aerodynamic ratio was positive, as expected. As speakers decreased their aerodynamic ratio by targeting a higher intraoral pressure, their productions tended to have a lower RFF. However, the $R^2$ values were low; the $R^2$ between Offset Cycle 10 RFF and the aerodynamic ratio and the $R^2$ between Onset Cycle 1 RFF and the aerodynamic ratio were .17 and .06, respectively. Conversely, when the $R^2$ was examined within individual speakers, the averaged within-speaker $R^2$ values were moderate. Specifically, the averaged within-speaker $R^2$ value between Offset Cycle 10 RFF and the aerodynamic ratio was .45. The averaged $R^2$ value between Onset Cycle 1 RFF and the aerodynamic ratio was .47. However, the $R^2$ values for individual speakers were highly variable. The $R^2$ values between offset cycle 10 RFF and the aerodynamic ratio ranged from .04 to .95. Similarly, the $R^2$ values between onset cycle 1 RFF and the aerodynamic ratio ranged from .05 to .87. The $R^2$ values between RFF and the aerodynamic ratio were low in productions made by some speakers (e.g., see top panel of Figure 3) yet high in productions made by other speakers (e.g., see bottom panel of Figure 3).

*Experiment 2.* During the /p/ sentences, offset RFF remained around 0 ST and onset RFF decreased as a function of cycle for the relaxed and typical conditions, whereas similar to Experiment 1, the RFF in the strained productions was generally lower (see Figure 2). When examined across speakers, the relationship between RFF and the perceptual measure was negative, as expected. That is, productions that were perceived as more effortful (higher perceptual score) had lower RFF. The $R^2$ between Offset Cycle 10 RFF and the perceptual ratings and the $R^2$ between Onset Cycle 1 RFF and the perceptual ratings were .21 and .26, respectively. When examined within individual speakers, the averaged $R^2$ between RFF and the perceptual ratings were moderate: The averaged $R^2$ values between offset cycle 10 RFF and perceptual ratings and between onset cycle 1 RFF and perceptual ratings were .46 and .56, respectively. Examining the $R^2$ values as a function of speaker revealed that the $R^2$ values between Offset Cycle 10 RFF and the perceptual ratings were highly variable, whereas the $R^2$ values between Onset Cycle 1 RFF and the perceptual ratings were mostly moderate to high. The $R^2$ values between Offset Cycle 10 RFF and the perceptual ratings of individual speakers ranged from < .01 to .94. The $R^2$ values between Onset Cycle 1 RFF and the perceptual ratings of individual speakers were all higher than .37 with the exception of one speaker. Again, the $R^2$ between RFF and the perceptual ratings were low in productions made by some speakers (e.g., see top panel of Figure 4) but high in productions made by other speakers (e.g., see bottom panel of Figure 4). The speakers with high $R^2$ values between RFF and perceptual ratings were

**Figure 4.** Offset Cycle 10 (left) and Onset Cycle 1 (right) RFF as a function of the perceptual ratings. Top: a participant whose productions yielded low $R^2$ between RFF and the perceptual ratings. Bottom: a participant whose productions yielded high $R^2$ between RFF and the perceptual ratings.



not consistently found to be the same speakers with high $R^2$ values between RFF and the aerodynamic ratio.

## Discussion

In order to determine its usefulness for tracking variations in effort in individual participants, we compared RFF with changes in an aerodynamic and a perceptual measure in participants who were creating purposeful modulations in their vocal effort. The RFF pattern for the relaxed, typical, and slightly strained conditions were similar to those observed in healthy young adults in previous studies (Robb & Smith, 2002; Watson, 1998); offset RFF remained around 0 ST for all cycles, and onset RFF decreased as a function of cycle. These observations suggest that speakers exhibited their typical RFF pattern even though they were slightly altering their vocal effort level. The RFF for the moderately strained and the maximally strained conditions exhibited a qualitatively different pattern from the one observed in the relaxed, typical, and slightly strained conditions. The RFF in these productions was generally lower, and the offset RFF decreased as a function of cycle, whereas the onset RFF was steady or even increased slightly as a function of cycle. This RFF pattern is similar to those observed in individuals with vocal hyperfunction (Stepp et al., 2010). Thus, when typical speakers drastically increased their vocal effort during voice production, this resulted in alterations to their RFF that mimicked the pattern observed in individuals with vocal hyperfunction.

Slight differences were noted in Onset Cycle 1 RFF between the /p/ sentences from Experiment 2 and the /pæ/ productions from Experiment 1, which were similar in nature to slight differences in RFF due to differences in stimuli choice that have been documented previously (Lien et al., 2014). Similarly, Smith and Robb (2013) also reported no significant difference in offset RFF taken from different phonetic contexts and a slight difference in onset RFF that depended more on laryngeal factors than on aerodynamic factors. Consequently, we did not expect the results of this study to depend on phonetic context. As expected, we found that RFF showed stronger relationships between the aerodynamic ratio and listener perception of vocal effort when examined within individuals relative to across individuals.

### *Relationships Within and Across Speakers*

RFF patterns differ between populations of hyperfunctional and typical speakers (Stepp et al., 2010). However, a previous study completed across individuals with voice disorders and differing levels of vocal hyperfunction found significant but weak correlations between RFF and listener perception of vocal effort (Stepp et al., 2012). The weak correlation suggests that the relationships among RFF, listener perception, and whether the speaker has vocal hyperfunction are not linear. Here we found relatively high correlations within speakers between both Offset Cycle 10 and Onset Cycle 1 RFF and listener ratings of vocal effort (average within-speaker $R^2$ = .46 and .56, respectively) relative to across-speaker relationships ($R^2$ = .21 and .26, respectively). It is challenging to draw conclusions about RFF and vocal effort across individuals due to the low overall variance explained by RFF across speakers. Nevertheless, the increase in correlations between listener perception and acoustic variables within speakers is similar to that shown in a previous study in which listeners were asked to estimate the distance between the speaker and the addressee based on the speaker's voice: all acoustic measures studied (e.g., sound pressure level, spectral emphasis, fundamental frequency) showed higher correlations with listener perceptions after correction for speaker-specific factors (Traunmuller & Eriksson, 2000).

Similarly, although no previous study has compared RFF with aerodynamic measures of vocal effort, previous work has compared aerodynamic and acoustic measures of voice within and across individuals (Holmberg, Hillman, Perkell, & Gress, 1994; Holmberg, Hillman, Perkell, Guiod, & Goldman, 1995). Relatively weak relationships between acoustic and aerodynamic measures are often found in group data, despite the fact that individual speakers show high correlations. In fact, Holmberg et al. (1994, p. 493) suggested that in voice patients, aerodynamic measures "seem more suitable for examination of individuals' changes from one level of vocal effort to another than for quantitative comparisons between patients and normal absolute values." The stronger relationships found between acoustic, aerodynamic, and perceptual measures within individuals relative to across individuals are likely due to differences between individual speakers in both their typical voice quality as well as in their technique of modulating vocal effort. In fact, speakers with high $R^2$ values between RFF and perceptual ratings were not consistently found to be the same speakers with high $R^2$ values between RFF and the aerodynamic ratio. This highlights the fact that vocal effort may be perceived differently by the speaker and the listener. Although this study examined healthy speakers who purposefully modulated their vocal effort, prior studies have shown that this finding also extends to individuals with dysphonia (Eadie et al., 2010; Lee et al., 2005). The difference between listener and speaker perception of vocal effort or overall severity may be due to differences in strategies used by listeners and speakers (Eadie et al., 2007, 2010; Lee et al., 2005).

Regardless of whether they were examined within or across speakers, correlations with listener perception of vocal effort were higher for onset cycle 1 RFF than for offset cycle 10 RFF. This corroborates previous findings that have suggested that whereas offset RFF may be well suited to detection of mild changes in vocal effort, onset RFF is more sensitive to differences in the level of vocal effort (Eadie & Stepp, 2013; Stepp et al., 2012). A simple model has been proposed to explain the physiological mechanisms behind RFF (Stepp et al., 2011; Watson, 1998). The model involves a combination of vocal fold kinematics, aerodynamics, and tension, incorporating the observation that the activity of the cricothyroid muscle tends to be high immediately preceding, during, and following the voiceless consonant (Löfqvist, Baer, McGarr, & Story, 1989). The increase in cricothyroid muscle activity can be associated with increases in tension and fundamental frequency (Löfqvist et al., 1989). Given that vocal hyperfunction is often associated with higher baseline tension (Hillman et al., 1989; Roy, Ford, & Bless, 1996), Stepp et al. (2011) hypothesized that the ability of individuals with vocal hyperfunction to use changes in tension to modulate their fundamental frequency is limited due to a ceiling effect, resulting in lowered RFF. However, this simple model does not account for differences between onset and offset RFF, suggesting that more research into the underlying mechanisms of changes in RFF is necessary.

### *Study Limitations*

The current study is limited due to the absence of a gold-standard measure of vocal hyperfunction or vocal effort with which to compare RFF. We compared the relationship between RFF and listener perception, which is limited by known issues with reliability in perceptual ratings of vocal effort or strain and which may be based on different cues of vocal effort than those self-rated by speakers (De Bodt et al., 1997; Wuyts et al., 1999). However, listener reliability issues appeared to be mitigated through our use of the sort and rate method (Granqvist, 2003) and the use of familiarization samples, resulting in strong intrarater and interrater reliability (average intrarater: $r = .93$; interrater: $\rho = .97$). The advantages of this procedure are like those shown by use of anchor samples or paired comparisons; however, it must be noted that the visual sort and rate method is not clinically viable due to the time necessary to complete the procedure. These reliability values found in this study are considerably higher than the ones observed in a previous study by Zraick et al. (2011) that evaluated the reliability of strain using the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V; American Speech-Language-Hearing Association, 2002) and the GRBAS scale: The average intrarater reliabilities for the CAPE-V and the GRBAS scale were $r = .35$ and $r_s = .53$, respectively; and the interrater reliabilities for the CAPE-V and the GRBAS scale were $\rho = .56$ and $\rho = .48$, respectively. Thus, use of the visual sort and rate method appears to hold promise for future experimental study.

We further compared RFF with the aerodynamic ratio of sound pressure level to subglottal pressure level using indirect estimates of subglottal pressure. Indirect measurement of subglottal pressure is difficult and can suffer from unreliable estimates. Furthermore, it is known to differ substantially from direct measures of subglottal pressure (measured invasively), particularly when individuals create changes in voice quality, such as in this study (McHenry, Minton, Kuna, Vanoye, & Roberts-Seibert, 1995). Future studies that incorporate direct measurements may be necessary to fully characterize the relationship between RFF and subglottal pressure. Incorporating direct measures of subglottal pressure are especially important when studying populations with voice disorders, particularly in populations with vocal hyperfunction. In these populations, indirect measures of subglottal pressure can be substantially more inaccurate due to increased medial (transglottal) pressure and stiffness of the vocal folds (Hillman et al., 1989). These physiological factors may create more separation between the subglottal and supraglottal space and thus may enlarge the difference between indirect measures (measured from the supraglottal space) and direct measures (measured from the subglottal space). Even with direct measures of subglottal pressure incorporated, it is still a challenge to estimate vocal effort based on aerodynamic measures alone. Although increases in vocal effort are often accompanied by high subglottal pressure and airflow, these are not the only factors that may contribute to the perception of vocal effort. Different strategies for creating vocal effort may explain the highly variable $R^2$ that were observed between RFF and the aerodynamic ratio in this study.

Finally, this study is limited by its use of individuals with healthy voices. These speakers were asked to create purposeful modulations in their vocal effort, which may or may not be an appropriate model for vocal hyperfunction in individuals with voice disorders: Individuals instructed to increase vocal effort may do so in a fundamentally different way than individuals with voice disorders related to vocal hyperfunction. Future work will compare these measures in individuals with vocal hyperfunction across the therapeutic process in order to document within-speaker changes in RFF in individuals with voice disorders.

## Conclusion

Speakers with healthy voices were asked to create purposeful modulations in their vocal effort, during which

RFF and both aerodynamic and perceptual measures of vocal effort were measured. During strained conditions, these participants displayed RFF patterns that were qualitatively different from their typical patterns and more similar to those observed previously in individuals with vocal hyperfunction. RFF showed stronger relationships between both the aerodynamic measure and listener perception of vocal effort when examined within individuals relative to across individuals. Future work is necessary to examine these relationships in individuals with vocal hyperfunction across the therapeutic process.

## Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grant DC012651 awarded to Cara E. Stepp. Thanks to Elizabeth Murray, Meredith Cler, Joseph Mendoza, and Felicia Patel for assistance with pilot vocal effort ratings and familiarization samples.

## References

**American Speech-Language-Hearing Association.** (2002). *Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): ASHA Special Interest Division 3, Voice and Voice Disorders*. Retrieved from http://www.asha.org/uploadedFiles/ASHA/SIG/03/affiliate/CAPE-V-Purpose-Applications.pdf

**Awan, S. N., & Roy, N.** (2006). Toward the development of an objective index of dysphonia severity: A four-factor acoustic model. *Clinical Linguistics & Phonetics, 20,* 35–49.

**Bhuta, T., Patrick, L., & Garnett, J. D.** (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice, 18,* 299–304.

**Boersma, W., & Weenink, D.** (2012). Praat: Doing Phonetics by Computer (Version 5.3.04) [Computer program]. Retrieved from http://www.praat.org/

**Brinca, L. F., Batista, A. P. F., Tavares, A. I., Gonçalves, I. C., & Moreno, M. L.** (2014). Use of cepstral analyses for differentiating normal from dysphonic voices: A comparative study of connected speech versus sustained vowel in European Portuguese female speakers. *Journal of Voice, 28,* 282–286.

**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,* 111.

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

**Eadie, T. L., Kapsner, M., Rosenzweig, J., Waugh, P., Hillel, A., & Merati, A.** (2010). The role of experience on judgments of dysphonia. *Journal of Voice, 24,* 564–573.

**Eadie, T. L., Nicolici, C., Baylor, C., Almand, K., Waugh, P., & Maronian, N.** (2007). Effect of experience on judgments of adductor spasmodic dysphonia. *Annals of Otology, Rhinology, & Laryngology, 116,* 695–701.

**Eadie, T. L., & Stepp, C. E.** (2013). Acoustic correlate of vocal effort in spasmodic dysphonia. *Annals of Otology, Rhinology, & Laryngology, 122,* 169–176.

**Faver, K. Y., Plexico, L. W., & Sandage, M. J.** (2012). Influence of syllable train length and performance end effects on estimation of phonation threshold pressure. *Journal of Voice, 26,* 18–23.

**Granqvist, S.** (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology, 28,* 109–116.

**Hertegård, S., Gauffin, J., & Lindestad, P.-Å.** (1995). A comparison of subglottal and intraoral pressure measurements during phonation. *Journal of Voice, 9,* 149–155.

**Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C.** (1989). Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech and Hearing Research, 32,* 373–392.

**Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C.** (1990). Phonatory function associated with hyperfunctionally related vocal fold lesions. *Journal of Voice, 4*(1), 52–63.

**Hillman, R. E., & Mehta, D. D.** (2011). Ambulatory monitoring of daily voice use. *SIG 3: Perspectives on Voice and Voice Disorders, 21,* 56–61.

**Hillman, R. E., Montgomery, W. W., & Zeitels, S. M.** (1997). Current diagnostics and office practice: Appropriate use of objective measures of vocal function in the multidisciplinary management of voice disorders. *Current Opinion in Otolaryngology—Head and Neck Surgery, 5*(3), 172–175.

**Hirano, M.** (1981). *Clinical examination of voice*. New York, NY: Springer Verlag.

**Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C.** (1994). Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings. *Journal of Speech and Hearing Research, 37*(3), 484–495.

**Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., & Goldman, S. L.** (1995). Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech and Hearing Research, 38,* 1212–1223.

**Ladefoged, P., & McKinney, N. P.** (1963). Loudness, sound pressure, and subglottal pressure in speech. *The Journal of the Acoustical Society of America, 35*(4), 454–460.

**Lee, M., Drinnan, M., & Carding, P.** (2005). The reliability and validity of patient self-rating of their own voice quality. *Clinical Otolaryngology, 30,* 357–361.

**Lien, Y.-A. S., Gattuccio, C. I., & Stepp, C. E.** (2014). Effects of phonetic context on relative fundamental frequency. *Journal of Speech, Language, and Hearing Research, 57,* 1259–1267.

**Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S.** (1989). The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America, 85,* 1314–1321.

**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 and Laryngology, 121,* 539–548.

**McHenry, M., Minton, J. T., Kuna, S. T., Vanoye, C. R., & Roberts-Seibert, N. S.** (1995). Comparison of direct and indirect calculations of laryngeal airway resistance in various voicing conditions. *European Journal of Disorders of Communication, 30,* 435–449.

**Netsell, R., Lotz, W., & Shaughnessy, A. L.** (1984). Laryngeal aerodynamics associated with selected voice disorders. *American Journal of Otolaryngology, 5*(6), 397–403.

**Plexico, L. W., Sandage, M. J., & Faver, K. Y.** (2011). Assessment of phonation threshold pressure: A critical review and clinical implications. *American Journal of Speech-Language Pathology, 20,* 348–366.

**Robb, M. P., & Smith, A. B.** (2002). Fundamental frequency onset and offset behavior: A comparative study of children and adults. *Journal of Speech, Language, and Hearing Research, 45,* 446–456.

**Rosenthal, A. L., Lowell, S. Y., & Colton, R. H.** (2014). Aerodynamic and acoustic features of vocal effort. *Journal of Voice, 28,* 144–153.

**Roy, N., Ford, C., & Bless, D.** (1996). Muscle tension dysphonia and spasmodic dysphonia: The role of manual laryngeal tension reduction in diagnosis and management. *Annals of Otology, Rhinology, and Laryngology, 105,* 851–856.

**Shrout, P. E., & Fleiss, J. L.** (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

**Smith, A. B., & Robb, M. P.** (2013). Factors underlying short-term fundamental frequency variation during vocal onset and offset. *Speech, Language, and Hearing, 16,* 208–214.

**Stepp, C. E., Hillman, R. E., & Heaton, J. T.** (2010). The impact of vocal hyperfunction on relative fundamental frequency during voicing offset and onset. *Journal of Speech, Language, and Hearing Research, 53,* 1220–1226.

**Stepp, C. E., Merchant, G. R., Heaton, J. T., & Hillman, R. E.** (2011). Effects of voice therapy on relative fundamental frequency during voicing offset and onset in patients with vocal hyperfunction. *Journal of Speech, Language, and Hearing Research, 54,* 1260–1266.

**Stepp, C. E., Sawin, D. E., & Eadie, T. L.** (2012). The relationship between perception of vocal effort and relative fundamental frequency during voicing offset and onset. *Journal of Speech, Language, and Hearing Research, 55,* 1887–1896.

**Titze, I. R.** (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America, 85,* 901–906.

**Traunmuller, H., & Eriksson, A.** (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America, 107,* 3438–3451.

**Verdolini, K., Titze, I. R., & Fennell, A.** (1994). Dependence of phonatory effort on hydration level. *Journal of Speech and Hearing Research, 37,* 1001–1007.

**Watson, B. C.** (1998). Fundamental frequency during phonetically governed devoicing in normal young and aged speakers. *The Journal of the Acoustical Society of America, 103,* 3642–3647.

**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,* 508–517.

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