PSYCHOMETRIC ANALYSIS OF AN ECOLOGICAL VOCAL EFFORT SCALE IN INDIVIDUALS WITH AND WITHOUT VOCAL HYPERFUNCTION DURING ACTIVITIES OF DAILY LIVING

Katherine L. Marks[1,2,*], Alessandra Verdi[1,2], Laura E. Toles[1,2], Kaila L. Stipancic[1,4], Andrew J. Ortiz[2,3], Robert E. Hillman[1,2,3], Daryush D. Mehta[1,2,3]

[1] MGH Institute of Health Professions, Boston, MA, USA
[2] Massachusetts General Hospital, Boston, MA, USA
[3] Harvard Medical School, Boston, MA, USA
[4] University at Buffalo, Buffalo, NY, USA

Objective: The purpose of this study was to examine the psychometric properties of an ecological vocal effort scale linked to a voicing task.

Methods: Thirty-eight patients with nodules, 18 patients with muscle tension dysphonia, and 45 vocally healthy control individuals participated in a week of ambulatory voice monitoring. A global vocal status question was asked hourly throughout the day. Participants produced a vowel-consonant-vowel-syllable string and rated the vocal effort needed to produce the task on a visual analog scale. Test-retest reliability was calculated for a subset using the Intraclass Correlation Coefficient (ICC(A,1)). Construct validity was assessed by: 1) comparing the week-long vocal effort ratings between the patient and control groups and 2) comparing week-long vocal effort ratings before and after voice rehabilitation in a subset of 25 patients. Cohen's *d*, the standard error of measurement (SEM), and the minimal detectable change (MDC) assessed sensitivity. The minimal clinically important difference (MCID) assessed responsiveness.

Results: Test-retest reliability was excellent (ICC(A,1): 0.96). Week-long mean effort was statistically higher in the patients than in controls (*d* = 1.62) and lower after voice rehabilitation (*d* = 1. 75), supporting construct validity and sensitivity. The SEM was 4.14, MDC was 11.47, and MCID was 9.74. Since the MCID was within the error of the measure, we must rely upon the MDC to detect real changes in ecological vocal effort.

Conclusion: The ecological vocal effort scale offers a reliable, valid, and sensitive method of monitoring vocal effort changes during the daily life of individuals with and without vocal hyperfunction.

Conflict of interest statement: Drs. Robert Hillman and Daryush Mehta have a financial interest in InnoVoyce LLC, a company focused on developing and commercializing technologies for the prevention, diagnosis and treatment of voice-related disorders. Dr. Hillman's and Dr. Mehta's interests were reviewed and are managed by Massachusetts General Hospital and Mass General Brigham in accordance with their conflict-of-interest policies.

Introduction

In the United States, voice disorders affect approximately one out of every thirteen adults annually, with far-reaching social, emotional, and economic consequences (Bhattacharyya, 2014). Vocal hyperfunction (VH), defined as excessive perilaryngeal musculoskeletal activity during phonation (Oates & Winkworth, 2008), is considered an etiological component in the most frequently occurring behavioral voice disorders (Bhattacharyya, 2014; Hillman et al., 1989; Hillman et al., 2020). One of the most frequent complaints of patients with VH is the requirement of increased vocal effort to speak (Colton et al., 2006; Hanschmann et al., 2011; Jiang & Titze, 1994; van Mersbergen et al., 2020). Vocal effort is defined as the perception of the work or exertion an individual feels during phonation (Hunter et al., 2020). This feeling can be particularly problematic for individuals who rely heavily on their voices throughout the day, such as teachers, singers, fitness instructors, lawyers, and clergy (Ramig & Verdolini, 1998; Roy et al., 2005; Roy et al., 2004; Verdolini & Ramig, 2001). It is no surprise then that reducing vocal effort is a frequent target in voice therapy (Hunter et al., 2020; van Mersbergen et al., 2020; Van Stan, Roy, et al., 2015). However, tracking and documenting degree of vocal effort in both clinical and research settings is challenging, as there is no standardized measure of vocal effort that is widely used (van Mersbergen et al., 2020).

Van Mersbergen et al. (2020) found that 78% of surveyed speech-language pathologists (SLPs) reported quantifying vocal effort using the Voice Handicap Index (VHI; Jacobson et al., 1989) or the shorter, 10-item VHI-10 (Rosen et al., 2004). This finding may be problematic because, although the VHI includes two items related to vocal effort, the VHI-10 does not; furthermore, these instruments were designed to measure the construct of voice disability, not vocal effort specifically. Other rating scales employed to quantify vocal effort include direct magnitude estimation scales (Banister, 1979; Tenenbaum et al., 2012; Verdolini et al., 1994), visual analog scales (VAS) (Borg, 1982; Borg, 1990; Gilman & Johns, 2017; McKenna & Stepp, 2018; Paes & Behlau, 2017; Shewmaker et al., 2010; Tanner et al., 2010), or Borg-derived scales, such as the OMNI vocal effort scale (Shoffel-Havakuk et al., 2019), the Borg CR-10 (Baldner et al., 2015; Borg, 1982; van Leer & van Mersbergen, 2017), and the Borg CR-100 (Berardi, 2020; Borg & Kaijser, 2006). These scales were intended to reflect patients' feeling of vocal effort at one point in time or their judged cumulative vocal effort. However, one-time ratings do not necessarily reflect patient reports of the ongoing changes in vocal effort they experience throughout the day—outside of the therapy session—that depend on their daily vocal demand, including environmental factors, number of communication partners, and type of voicing activity (Hunter et al., 2020; Van Stan, Maffei, et al., 2017). Moreover, the in-clinic ratings do not provide insight to the physiological underpinnings that may accompany changes in vocal effort throughout daily activities.

Vocal effort measured in daily life should better reflect the changes that occur throughout the day, depending on vocal demands. This information can inform the treating speech-language pathologist's therapy strategies and help facilitate the carryover of such strategies from the clinic to voice use during activities of daily living. Further, pairing real-world effort judgments with objective measures from ambulatory voice monitoring (Mehta et al., 2015) should provide better insights into the physiology underlying daily variation in vocal effort, which could then be implemented as an early-warning system to alert individuals when they begin to exhibit vocal behaviors that could influence their vocal effort. The measures could also be employed for biofeedback to aid in behavioral self-regulation for patients with voice disorders (Llico et al., 2015; Van Stan et al., 2014; Van Stan, Mehta, et al., 2015; Van Stan, Mehta, Petit, et al., 2017; Van Stan, Mehta, Sternad, et al., 2017). These objective measures could also be used to document changes throughout daily life induced by voice therapy. Identifying objective measures underlying changes in vocal effort throughout an individual's day has the potential to change the way SLPs assess, treat, and ultimately prevent behavioral voice disorders.

Ecological Momentary Assessment of Vocal Status

Ecological momentary assessment (EMA) involves assessing individuals' current experiences and/or behaviors as they occur in real time and in their real-world setting (Burke et al., 2017). Advantages of EMA include prompting individuals to rate or answer questions "in the moment" to minimize recall bias, obtain self-ratings in their natural environment as opposed to controlled laboratory conditions or clinical situations, and correlate these ratings to underlying physiological processes (Burke et al., 2017; Shiffman et al., 2008). Because a person's vocal effort may be affected by vocal demands that vary throughout their day, it is likely that there will be large variability in self-ratings throughout a day or week. Ecological measurement of voice is not a new concept in the voice literature; self-reports of vocal status (e.g., vocal fatigue, discomfort, difficulty to produce soft phonation, and vocal effort) have been collected multiple times throughout a day or before and after heavy voice use in vocally healthy individuals and occupational voice users with presumably healthy voices (Gotaas & Starr, 1993; Kitch et al., 1996; Lehto et al., 2008; Vintturi et al., 2003; Welham & Maclagan, 2004). The development of ambulatory voice monitoring (Bottalico et al., 2018; Cheyne et al., 2003; Mehta et al., 2015; Mehta et al., 2012; Popolo et al., 2005; Van Stan et al., 2014) has made possible the ability to capture changes in voicing and elicit self-ratings of vocal status during activities of daily life (Carroll et al., 2006; Dallaston & Rumbach, 2016; Halpern et al., 2009; Hunter & Titze, 2009; Hunter & Titze, 2010; Laukkanen et al., 2008; Laukkanen & Kankare, 2006; Popolo et al., 2011; Van Stan, Maffei, et al., 2017).

Several studies have examined ecological voice ratings using ambulatory voice monitoring systems, but only in individuals with healthy voices (Carroll et al., 2006; Halpern et al., 2009; Hunter & Titze, 2008; Van Stan, Maffei, et al., 2017; Verdyuckt et al., 2011). Carroll and colleagues (2006) first used ambulatory voice monitoring to capture ratings of the inability to produce soft voice (IPSV) during low intensity tasks and ratings of vocal effort during loud phonation tasks in seven vocally healthy male singers, using a personal digital assistant. Since then, other work has been done to investigate ecological momentary assessment of vocal status in relation to vocal dose measures (Halpern et al., 2009; Hunter & Titze, 2008; Verdyuckt et al., 2011). However, these studies have yielded limited success using traditional ambulatory measures related to pitch, loudness, and vocal doses, to quantify changes in self-reported vocal status.

More recently, Lei et al. (2020) used a vocal dose–based vocal loading task to investigate the relationship between voice use and vocal fatigue in ten vocally healthy participants, who participated in six consecutive 30-minute vocal loading tasks. They found that vocal effort and discomfort scores increased rapidly between the first and second loading tasks, whereas the IPSV score increased to a lesser degree. This finding suggests that participants may perceive effort and discomfort even when their vocal demand response (i.e., IPSV task) is less affected. The acoustic features related to distance dose (i.e., fundamental frequency [f_o], sound pressure level, percent phonation) followed the same trend of vocal effort and discomfort scores, with a sharp increase in the early vocal loading tasks that remained steady through the rest of the vocal loading tasks. The authors did not, however, look specifically at the relationship between the acoustic measures and the self-ratings.

Van Stan et al. (2017) was the first study, to our knowledge, to measure patient-perceived vocal status throughout daily life. They validated self-ratings of vocal status in individuals with and without vocal hyperfunction using a smartphone-based ambulatory voice monitoring system to prompt participants to rate difficulty to produce soft high-pitched phonation (similar to IPSV), vocal discomfort, and vocal fatigue using a VAS periodically every 5 hours throughout the day. The study provided evidence of reliability and validity for tracking vocal status in daily life. The authors found internal consistency among the three questions, reflecting the construct of vocal status. They found a minimally detectable change (MDC) using a 95% confidence interval (MDC_{95}) of approximately 20 points for each vocal status dimension, indicating a true change is detectable when participants change their vocal status ratings by 20 points or more. The study demonstrated known-groups validity by determining statistically significant differences in mean self-ratings between individuals with and without vocal hyperfunction, as well as statistically significant differences in mean vocal status self-ratings for individuals with VH before and after successful voice treatment (i.e., therapy and/or surgery) (Van Stan, Maffei, et al., 2017). Although accelerometer-based ambulatory voice measures were not specifically investigated in that study, the dataset lends itself to the study of ambulatory voice measures associated with changes in vocal status in patients with vocal hyperfunction.

Ecological Momentary Assessment of Vocal Effort in Patients with VH

Although Van Stan and colleagues (2017) provided an empirically validated set of vocal status questions used in ambulatory voice monitoring, the authors did not specifically investigate ratings of vocal effort. Other than anecdotal patient reports, little is known about how patients' perception of vocal effort changes throughout a week, depending on their specific vocal demands. The overarching aim of the current study was to examine the ecological momentary assessment of vocal effort, defined as the amount of perceived work or exertion to produce voice measured in an individual's real-world speaking environment. Specifically, we examined the psychometric properties (reliability, validity, sensitivity, and responsiveness) of an ecological vocal effort scale that was temporally linked to a voicing task and used to capture vocal effort ratings throughout a week of ambulatory voice monitoring in individuals with and without VH, with the ultimate the goal of providing a generalizable method to measure vocal effort throughout daily life.

Method

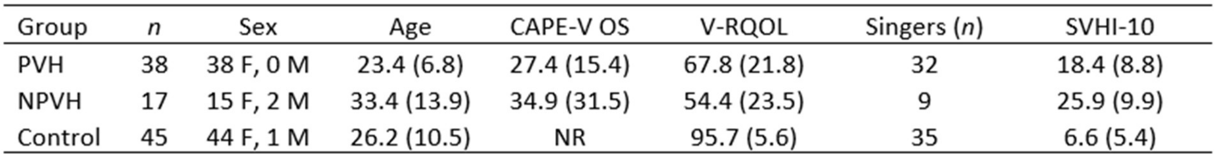
Participants

Hillman et al. (2020) differentiated two types of vocal hyperfunction: 1) phonotraumatic vocal hyperfunction (PVH) which includes benign vocal fold lesions (e.g., nodules) and 2) nonphonotraumatic vocal hyperfunction (NPVH), defined as dysphonia that occurs in the absence of concurrent known pathology (i.e., primary muscle tension dysphonia). Patients with either PVH or NPVH were recruited to the study via convenience sampling from the Center for Laryngeal Surgery and Voice Rehabilitation at the Massachusetts General Hospital (MGH Voice Center). Diagnosis was based on a comprehensive team evaluation by a laryngologist and SLP, including a complete case history, videostroboscopic evaluation, acoustic and aerodynamic assessment, patient-reported voice-related quality of life (V-RQOL) questionnaire (Hogikyan & Sethuraman, 1999), and SLP-rated consensus auditory-perceptual evaluation of voice (CAPE-V; Kempster et al., 2009). During this team evaluation, there is a patient-centered discussion between the patient and the clinicians regarding the history, voice use, diagnosis, and treatment options (e.g., therapy, surgery, or a combination of both). Generally, but not always, voice therapy is suggested as the first treatment approach at the MGH Voice Center (Van Stan, Mehta, Ortiz, Burns, Marks, et al., 2020). Ultimately, it is the patient who decides the course of treatment after discussing options and recommendations from the team.

Control participants without VH were recruited via snowball sampling: enrolled patients were asked to identify colleagues who matched their age (± 5 years), sex, and occupation, as well as singing genre (if a professional singer) as part of a larger ongoing study (Mehta et al., 2015). At the time of this group-based project, not all participants had been matched. Control participants were screened by a voice-specialized SLP to ensure 1) typical hearing in both ears through pure-tone air conduction at 25 dB HL at 0.5, 1, 2, and 4 kHz, 2) typical sounding voice, and 3) straight vocal fold edges with typical vibration patterns as observed via videostroboscopic examination.

Thirty-eight patients with phonotraumatic vocal hyperfunction (PVH), 17 patients with non-phonotraumatic vocal hyperfunction (NPVH), and 45 control individuals without vocal hyperfunction (VH) were enrolled as part of a larger ongoing study (Mehta et al., 2015). Table 1 lists descriptive data for all participants with respect to age, sex, overall severity (OS) from the CAPE-V, V-RQOL, and Singing Voice Handicap Index-10 (SVHI-10) for singers. A majority of participants were students studying voice at the collegiate level; these student singers made up 61% of the PVH group, 39% of the NPVH group, and 64% of the control group. Singers were only included in the NPVH group if speaking voice use was negatively impacted by the disorder. For those who were not student singers, the occupations varied across participants. (Cohen et al., 2009). Figure 1 illustrates a flowchart of participants through the different phases and analyses of the study.

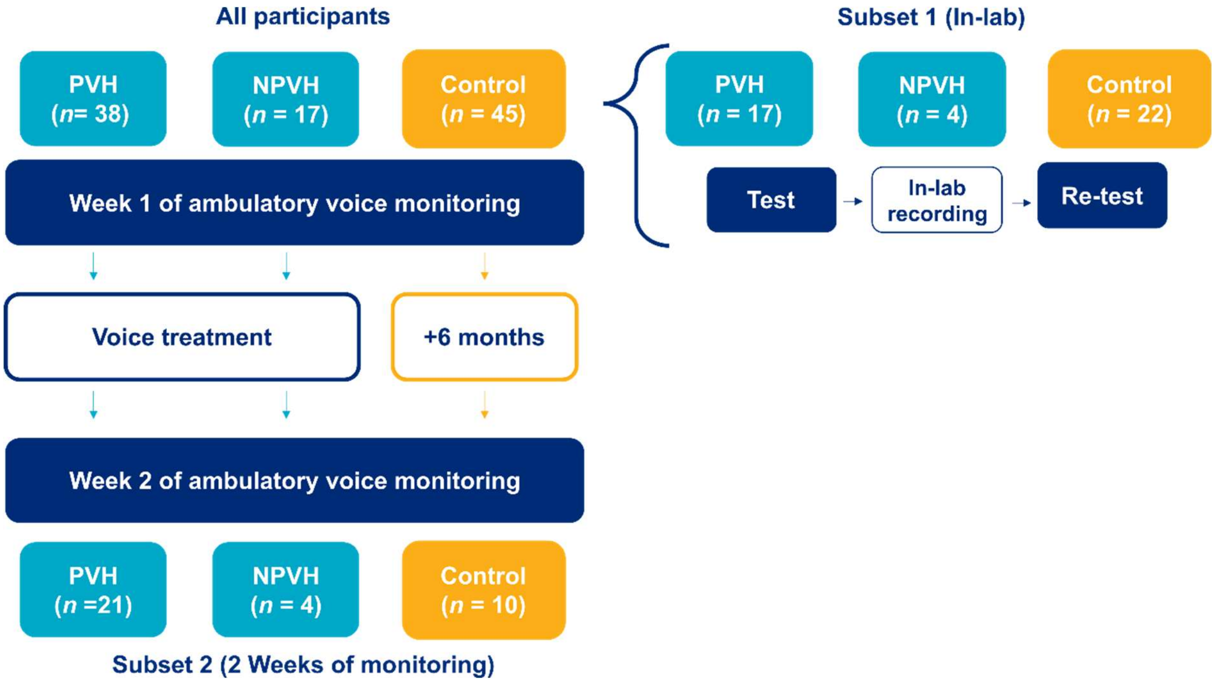
Table 1. Characteristics of participants by group. The phonotraumatic vocal hyperfunction (PVH) group consisted of 38 females, the non-phonotraumatic vocal hyperfunction (NPVH) group consisted of 15 females and two males, and the control group consisted of 44 females and 1 male. Gender information was not collected. Mean (standard deviation) age, overall severity from the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V), Voice-Related Quality of Life (V-RQOL) scores, number of singers in each group, and Singing Voice Handicap Index-10 (SVHI-10) for those singers are described. NR = not rated.



Group	*n*	Sex	Age	CAPE-V OS	V-RQOL	Singers (*n*)	SVHI-10
PVH	38	38 F, 0 M	23.4 (6.8)	27.4 (15.4)	67.8 (21.8)	32	18.4 (8.8)
NPVH	17	15 F, 2 M	33.4 (13.9)	34.9 (31.5)	54.4 (23.5)	9	25.9 (9.9)
Control	45	44 F, 1 M	26.2 (10.5)	NR	95.7 (5.6)	35	6.6 (5.4)

Figure 1. Flowchart illustrating methods and breakdown of each subset: All participants completed one week of ambulatory voice monitoring prior to any voice treatment. Participants in Subset 1 rated vocal effort before and after a voice recording that took place in a laboratory environment. These ratings were used for the test-retest analysis. Subset 2 included participants who completed a second week of monitoring. The patient participants completed their second week of ambulatory voice monitoring

following discharge from voice therapy, and the control participants completed a second week of monitoring at least six months after their initial week of monitoring.



Subset 1

Two subsets of participants were used in the study, as outlined in Figure 1. The first subset, displayed in Figure 1 (Subset 1) was used in a test-retest reliability analysis and included 14 female participants with PVH, 5 female and 2 male participants with NPVH, and 22 female control participants who were enrolled in the last year of the study. A test-retest protocol was only in place the final year of the project, which limited the number of participants included in the test-retest analysis. Table 2 describes Subset 1, listing the phases of the study each participant was in during the test-retest protocol. The average CAPE-V overall severity (OS) score was in the mild range for both patient groups. Specifically, the OS score was 17.0 (SD = 8.8) for patients with PVH and 28.7 (SD = 28.3) for patients with NPVH.

Table 2. Number of participants by group (phonotraumatic vocal hyperfunction [PVH], non-phonotraumatic vocal hyperfunction [NPVH], and controls) in Subset 1 whose scores were used in the test-retest analysis, including which phase of the study the test-retest protocol took place. Participants in the PVH and NPVH group participated either before treatment (Pre-Tx), after successful treatment (Post-

Tx), or follow-up at least six months later. Participants in the control group participated in either a baseline session or follow-up at least six months later. Em dash (—) indicates not applicable.

Subset 1 Group	Pre-Tx/Baseline	Post-Tx	6-mo. Follow-up
PVH (n = 14)	1	4	9
NPVH (n = 7)	4	1	2
Controls	15	—	7

Subset 2

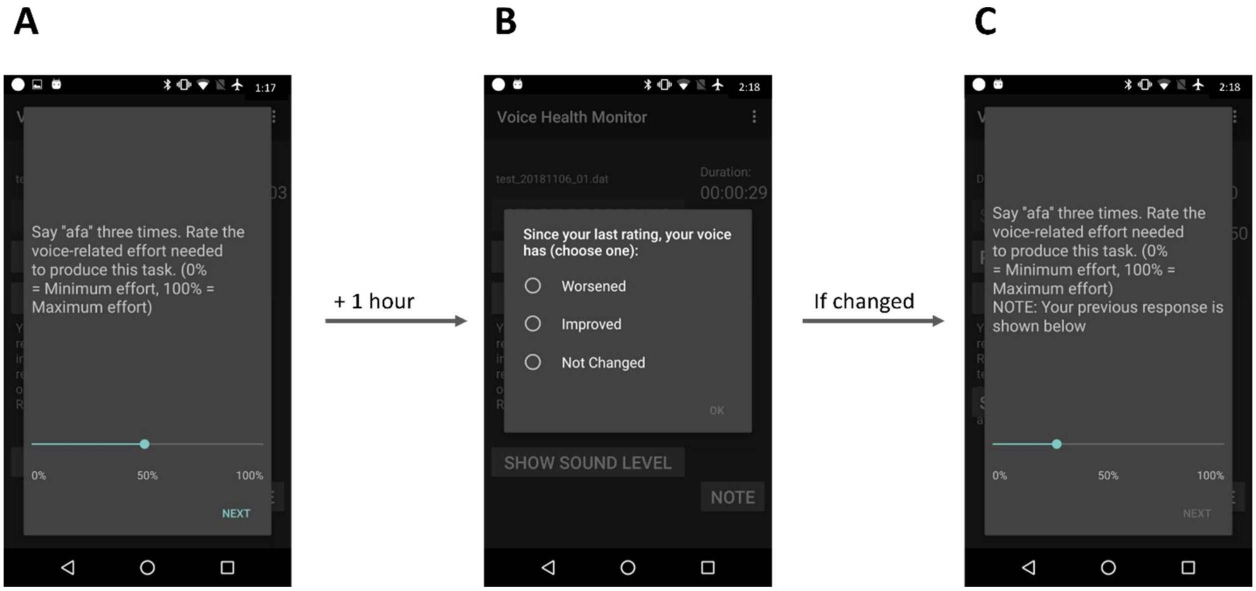
Subset 2, displayed in Figure 1, included 25 patients with VH (21 females with PVH; 2 females, 2 males with NPVH) who participated in multiple weeks of voice monitoring. The 19 patients who received voice therapy participated in a second week of monitoring after they completed a full course of voice therapy and were officially discharged from therapy following a comprehensive voice evaluation with subjective judgments of improvement from both the patient and the treating SLP. Four patients with PVH had surgery to remove phonotraumatic lesions and participated after they were discharged from post-operative voice therapy. Ten participants in the control group participated in a second week of voice monitoring at least six months after their initial baseline week of monitoring. Though not used in any of the analyses other than the test-retest protocol, patients also participated in a follow-up week of monitoring six months after voice rehabilitation. Table 3 describes participant characteristics for Subset 2, including age, sex, overall severity (OS) from the CAPE-V, V-RQOL, and Singing Voice Handicap Index-10 (SVHI-10) for singers.

Procedures

When participants were enrolled in the study, they participated in a one-hour lab visit, during which they were taught how to use the ambulatory monitoring equipment (smartphone and neck surface accelerometer sensor; Mehta et al., 2015) and how to respond to the vocal effort and global vocal status prompts, as described in the following paragraph. As part of a larger study, participants engaged in an in-lab recording session, which took approximately 20 minutes of the one-hour visit. Participants then took the equipment home with them to wear during their waking hours for approximately seven days. During the final year of data collection, the researchers implemented a test-retest protocol, in which they asked a subset (Subset 1) of participants to rate their vocal effort before and after the ~20-minute recording session. Although the session took approximately 20 minutes, the recording itself was between 6 to 8 minutes and the participants were asked to repeat vowel sounds, give a 30-second spontaneous speech sample, and read a passage and a list of sentences out loud. The recording was not expected to impact the ratings of participants. Ratings before and after the recording session were used for a test-retest analysis, further described in the "Psychometric Analyses" section. During the COVID-19 pandemic, the "in lab" visit was converted to a virtual visit via a HIPAA-compliant videoconferencing portal. The test-retest protocol remained the same, as researchers asked the participants to rate vocal effort before and after the ~20-minute recording session.

During their week of ambulatory voice monitoring, each morning when participants pressed the "Start recording," button on the smartphone platform, they were asked to rate their perceived vocal effort. Vocal effort was described using its definition (Hunter et al., 2020); specifically, the clinician investigators explained to participants that vocal effort refers to how much work it takes to speak or sing in the moment. Figure 2A displays a screenshot of the vocal effort prompt: "Say afa three times. Rate the voice-related vocal effort needed to produce this task." The syllable string "afa" was selected to allow for future analysis of relative fundamental frequency, since relative fundamental frequency has been theoretically and empirically associated with vocal effort (Lien et al., 2015; McKenna et al., 2016; Stepp et al., 2010; Stepp et al., 2011). Vocal effort ratings were made using a VAS on the smartphone in portrait mode, consistent with prior work by Van Stan, Maffei, et al. (2017), with labels of 0% at the left end of the scale for minimum effort, 50% in the center of the scale, and 100% on the right end of the scale for maximum effort. The cursor on the VAS was defaulted to the center of the scale each morning, and participants were required to move the cursor to indicate their current level of voice-related effort. After the initial prompting each morning, participants were alerted at hourly intervals to indicate whether their overall (i.e., global) vocal status had changed since their last rating, as shown in Figure 2B. If participants indicated that their vocal status had worsened or improved, they were prompted to re-rate their vocal effort, and their most recent ratings were displayed so that their responses were anchored to their previous rating, as displayed in in Figure 2C. If participants reported that their vocal status had not changed, they were not asked to re-rate. Participants were always required to answer the vocal effort prompt at the beginning of the recording each day and the end of the recording each day. Note that ratings for voice-related discomfort and fatigue were similarly collected on a VAS just prior to the vocal effort rating. This study focused solely on the third prompt of vocal effort, with the global vocal status used as an indicator of overall change.

Figure 2. Flowchart of smartphone screenshots of vocal status prompts. Each morning, when participants started recording, they were shown the vocal effort prompt and asked to say "afa" three times and rate the voice-related effort needed to produce the task (A). The cursor started at 50% and participants moved the cursor on the visual analog scale. Each hour, the phone vibrated to alert participants that it was time to answer a global vocal status question (B), where they are prompted to answer whether their voice had worsened, improved, or not changed since their last rating. If they selected "No Changed" on this global status prompt, no follow-up prompts were displayed until the next hour. If they selected "Worsened" or "Improved," they were asked to re-rate their voice-related effort (C). The cursor was displayed on the scale where they last provided an effort rating, so that their current rating was anchored to this previous rating. Once they produced the "afa" task and re-rated their vocal effort to produce that task, they were done until the next hour, when the global status prompt reoccurred.



Say "afa" three times. Rate the voice-related effort needed to produce this task. (0% = Minimum effort, 100% = Maximum effort)

Since your last rating, your voice has (choose one):

○ Worsened

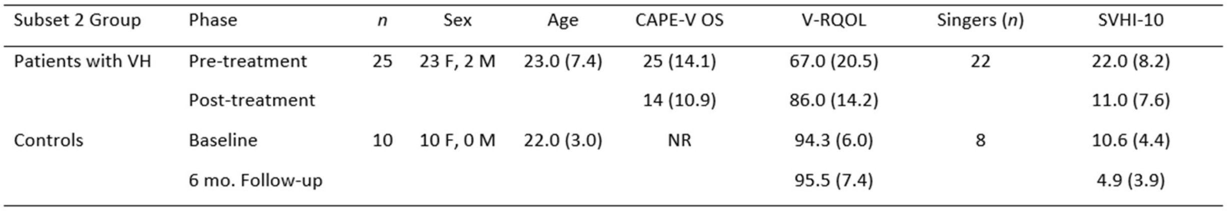
○ Improved

○ Not Changed

Say "afa" three times. Rate the voice-related effort needed to produce this task. (0% = Minimum effort, 100% = Maximum effort) NOTE: Your previous response is shown below

Participants in Subset 2 (Table 3; Figure 1) included those patients who completed follow-up weeks of ambulatory monitoring after successful voice rehabilitation (i.e., voice therapy and/or surgery or voice therapy only) or, for the control group, a follow-up week at least six months after the initial week of monitoring. The same procedures described were implemented for follow-up weeks of ambulatory voice monitoring. Although not used in the analyses other than test-retest reliability, patients also participated in a follow-up at least six months after successful treatment.

Table 3. Subset 2 includes 25 patients with vocal hyperfunction (VH) who participated before and after successful treatment and 10 individuals without VH in the control group who participated for a baseline week of monitoring and a follow-up week of monitoring at least six months later. Mean and standard deviation (SD) displayed for Age, Overall Severity (OS) from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), Voice-Related Quality of Life scores (V-RQOL), number of singers in each group, and Singing Voice Handicap Index-10 (SVHI-10) for those singers.



Subset 2 Group	Phase	*n*	Sex	Age	CAPE-V OS	V-RQOL	Singers (*n*)	SVHI-10
Patients with VH	Pre-treatment	25	23 F, 2 M	23.0 (7.4)	25 (14.1)	67.0 (20.5)	22	22.0 (8.2)
	Post-treatment				14 (10.9)	86.0 (14.2)		11.0 (7.6)
Controls	Baseline	10	10 F, 0 M	22.0 (3.0)	NR	94.3 (6.0)	8	10.6 (4.4)
	6 mo. Follow-up					95.5 (7.4)		4.9 (3.9)

Psychometric Analysis

The ratings of vocal effort and ratings of global vocal status for each time point were extracted from a smartphone file that maintained a timestamped log of user interactions and input for each question. Data were cleaned to remove irrelevant or repeated ratings that occurred during the in-lab visit, except for those used for test-retest reliability. Psychometric analyses were performed to assess test-retest reliability, construct validity, sensitivity to change, and responsiveness, for the ecological vocal effort scale.

Test-Retest Reliability

Reliability reflects the amount of both random and systematic error inherent in any measurement (Streiner et al., 2015). The intra-class correlation (ICC) is a measure of reliability that is defined as a ratio of participant variability over the product of participant variability and measurement error. The reliability coefficient expresses the proportion of the total variance in the measurements that is due to “true” differences between participants (Streiner et al., 2015). Historically, there has been a lack of consistent approaches regarding which ICC formula is appropriate for test-retest reliability in patient-reported outcomes and a lack of a uniform naming convention for the ICC formulas. Specifically, a key limitation in the general ICC literature is the use of the term “raters,” which does not easily translate to patient-reported outcomes, which typically involve the same raters evaluated at two different time points (Qin et al., 2019). Thus, the Critical Path Institute’s Patient-Reported Outcome (PRO) Consortium performed an extensive review of the literature on ICCs and presented their recommendations to be vetted by a group of 12 experts, including psychometricians, biostatisticians, regulators, and other scientists representing the PRO Consortium, the pharmaceutical industry, clinical research organizations, and consulting firms (Coons et al., 2011). To assess test-retest reliability for PRO measures, Coons and colleagues (2011) recommend using a two-way mixed-effect analysis of variance (ANOVA) model (fixed effect of two test periods and random effect of rater), with interaction for the absolute agreement between single scores, which is ICC(A,1) (Qin et al., 2019) .

Test-retest reliability was performed on the data from Subset 1 (Table 2; Figure 1), which included participants who rated vocal effort before and after a recording session. ICC(A,1) was used to obtain a correlation between ratings before and after the session. Because reliability metrics are not in the same units as the measure of interest, reliability estimates should be accompanied by the standard error of the measurement (SEM), which is expressed in the same unit of measurement as the original scores (Streiner et al., 2015). In the present study, the SEM was calculated using the equation $\text{SEM} = \sigma\sqrt{1-R}$, where σ is the standard deviation of the observed scores from the entire dataset of patients and controls during the initial week of monitoring, and R is the reliability coefficient ICC(A,1). We also calculated the test-retest reliability for two groups within Subset 1, employing a data-driven approach that uses overall severity (OS) of dysphonia, as rated by a voice-specialized SLP, for two groups: those judged by an SLP as within functional limits (WFL), defined as ≤ 10 on the CAPE-V, and those with mild overall severity, defined as

> 10 and ≤ 35 on the CAPE-V (Solomon et al., 2011). There were not enough participants with moderate or severe scores on the CAPE-V in Subset 1.

Validity

In general terms, validation of a scale involves determining a degree of confidence that can be placed on the inferences made about people based on their scores from the scale (Landy, 1986). Historically, validity has been divided into content, criterion, and construct validity; in recent years, the focus has shifted more to the logic and methodology of hypothesis testing (Streiner et al., 2015). With respect to both approaches, in the current study, we evaluated construct validity based on two hypotheses: 1) Individuals with VH will report higher week-long mean ecological vocal effort than individuals without VH, and 2) Individuals with VH will have lower week-long mean ecological vocal effort after successful treatment, compared to their pre-treatment week-long ecological vocal effort. To test the first hypothesis, a one-way ANOVA was used to test the difference in week-long mean vocal effort among the three groups (PVH, NPVH, and controls). Welch's *F* test was used, as the groups had unequal variances. To test the second hypothesis, a paired-samples *t*-test was used to assess differences in week-long mean vocal effort before and after successful voice treatment using data from Subset 2.

Sensitivity to Change

Sensitivity to change reflects an instrument's ability to measure any degree of change, regardless of whether it is relevant or meaningful to the decision maker (Liang, 2000; Streiner et al., 2015). The most well-known of sensitivity measures is Cohen's *d*, (Cohen, 1988), which is the mean ratio of the mean difference to the standard deviation (SD) of baseline scores (Streiner et al., 2015). Two analyses were performed, first comparing patients with VH and controls and second comparing patients before and after voice rehabilitation.

The minimal detectable change (MDC) is a commonly reported reference for interpretation of clinical outcome measures (Stipancic et al., 2018; Tilson et al., 2010; Van Stan, Maffei, et al., 2017). The MDC is defined as the smallest amount of change that is greater than measurement error (Beckerman et al., 2001; Haley & Fragala-Pinkham, 2006). The MDC with 95% confidence intervals (MDC_{95}) was used in this study as one index of responsiveness, and was calculated using the formula $\mathrm{MDC}_{95} = \mathrm{SEM}(1.96)\sqrt{2}$, with 1.96 representing the z-score for a 95% confidence interval and the $\sqrt{2}$ accounting for the difference of the two variances used to derive the SEM (Tilson et al., 2010). Although the MDC indicates that a change detected is unlikely due to chance variability, the MDC does not indicate whether or not the degree of change is clinically meaningful (Beninato & Portney, 2011). Thus, the minimal clinically important difference (MCID) was also used in the current study as an index of responsiveness.

Responsiveness

Responsiveness is the ability of an instrument to measure a meaningful or clinically important change in a clinical state (Liang, 2000). This change can be from the perspective of a patient, a caregiver, or a health professional. Although commonly studied in the physical and occupational therapy literature, very few studies have investigated responsiveness of voice and speech outcomes (Stipancic et al., 2018; Van Stan, Maffei, et al., 2017). Application of responsiveness indices are critically important for learning about how vocal effort changes throughout the day and assessing when treatments are making real and clinically important changes for patients with VH. To calculate the MCID, an anchor-based approach was employed using the vocal status ratings of “worsened,” “improved,” or “not changed,” to evaluate participants’ perception of overall change (Jaeschke et al., 1989). The change in (delta) effort score for each repeated rating was made, subtracting rating 2 from rating 1 and so on, for all participants during their first week of ambulatory voice monitoring. When participants reported that their vocal status had “Not Changed,” the delta effort score was assumed to be 0 (as participants were not asked to rate their vocal effort at that time). This assumption introduces an intentional bias implemented to be less bothersome for participants probed throughout the day. Moreover, we contend that only participants can offer the “ground truth” for themselves; so, when they say no change has occurred, we must assume that no change occurred. The MCID for changes in vocal effort was calculated as the average of the absolute delta effort scores for ratings following “worsened” or “improved” vocal status.

To further explore the data, we also stratified participants by OS and calculated MDC_{95} and MCID for each overall severity level, consistent with our ICC methods. It should be noted that because there were few participants in moderate and severe groups, we were unable to calculate specific ICCs for those groups, so we used the overall ICC to calculate the SEM for the moderate and severe groups. Thus, the MDC and MCID results for the moderate and severe groups were "best-case scenario" results for a small number of participants.

Results

Test-Retest Reliability

The overall ICC(A,1) was 0.96, indicating excellent test-retest reliability. The SEM, in the same units as ecological vocal effort, was found to be 4.14. The SEM, an absolute measure, quantifies the precision of scores within the participants. When stratified by overall severity of dysphonia, the ICC(A,1) for the WFL OS group was 0.87, indicating good test-retest reliability. The SEM for the WFL OS group was 1.95. For the Mild OS group, the ICC(A,1) was 0.91, indicating excellent reliability. The SEM for the mild OS group was 5.35. There were not enough participants in the moderate or severe groups to calculate specific ICCs, so the overall ICC was used to calculate specific SEMs: 5.45 for the moderate OS group and 5.27 for the severe OS group.

Validity

Two hypotheses were tested to establish construct validity. Levene's test indicated that the groups (PVH, NPVH, and Controls) had unequal variances, violating the assumption of homogeneity ($F(2,62) = 17.39$, $p < .001$). Therefore, Welch's F was used to test the differences in week-long mean vocal effort among groups, revealing a statistically significant main effect of diagnosis on week-long mean vocal effort scores ($F(2,19) = 13.44$, $p < .001$), with a medium effect size ($\eta^2 = .59$). Bonferroni-corrected pairwise comparisons, which divided the α level of significance of .05 by the number of tests performed (3), revealed statistically significant differences**,** with large effect sizes, between the PVH group and the controls ($p < .01$, $d = 1.62$) and between the NPVH group and the controls ($p < .01$, $d = 1.61$). Figure 3 illustrates week-long mean vocal effort for each group, with error bars indicating group-wide standard deviations (SD), and shading illustrating the range of mean vocal effort. Table 4 reports the week-long mean vocal effort statistics for each participant group.

Figure 3. Group-wide statistics (mean, standard deviation, range) for week-long mean vocal effort, demonstrating construct validity between individuals with vocal hyperfunction and vocally healthy controls. Dark blue bars display week-long mean vocal effort across individuals in each group. Error bars indicate standard deviation. Light blue shading indicates the range of week-long mean vocal effort scores.

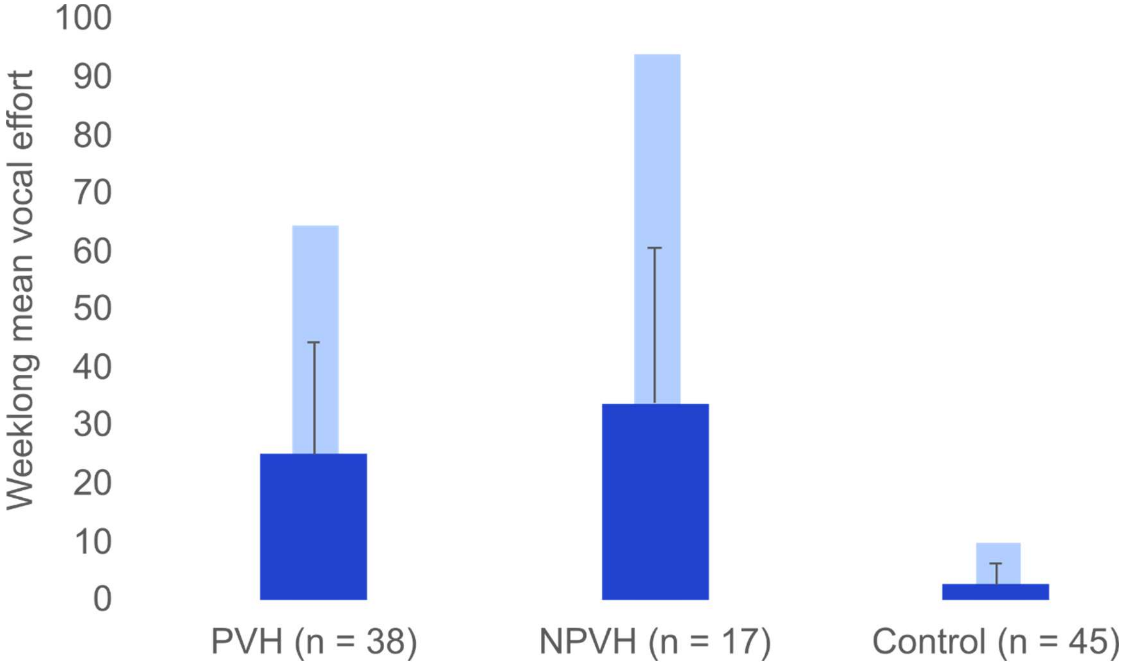


Table 4. Group mean, standard deviation (SD), minimum (Min), and maximum (Max) of the week-long average of vocal effort scores for participants during their first week of voice monitoring (pre-treatment for patients and baseline for controls).

Group	*n*	Group mean (SD) of mean week-long vocal effort	Min–Max
PVH	38	25.0 (19.4)	0.0–65.7
NPVH	17	33.8 (26.9)	1.3–94.2
Controls	45	2.6 (3.6)	0.0–12.2

Because week-long vocal effort was not different between the two patient groups (PVH and NPVH), the patient groups were collapsed into a single vocal hyperfunction group for the second validity analysis, which compared week-long mean vocal effort before and after successful treatment. To address the second hypothesis, the paired-samples *t*-test revealed that week-long mean vocal effort was statistically lower after successful voice rehabilitation (i.e., voice therapy and/or surgery), with a very large effect size (*t*(24) = 4.33, *p* < .001, *d* = 1.77). The mean of the differences in vocal effort was 14 points. This finding provides secondary evidence of construct validity for the ecological vocal effort scale. Complementary to the second hypothesis, a subset of controls was monitored for a second week at least six months after their initial baseline week. As expected, week-long vocal effort was relatively stable from baseline to six-month follow-up in this control group (*p* = .22). Figure 4 compares the week-long mean vocal effort statistics pooling all patients with VH before and after successful treatment and for controls at baseline and follow-up time points. Table 5 reports the week-long mean vocal effort statistics displayed in Figure 4, in addition to statistics separately for the patient groups with PVH and NPVH.

Figure 4. Group-wide statistics for week-long mean vocal effort before and after treatment for patients with vocal hyperfunction (VH) displayed in teal and baseline and follow-up of at least six months for controls displayed in yellow, demonstrating treatment-related construct validity. Error bars indicate standard deviations, and shading indicates range of week-long mean vocal effort scores.

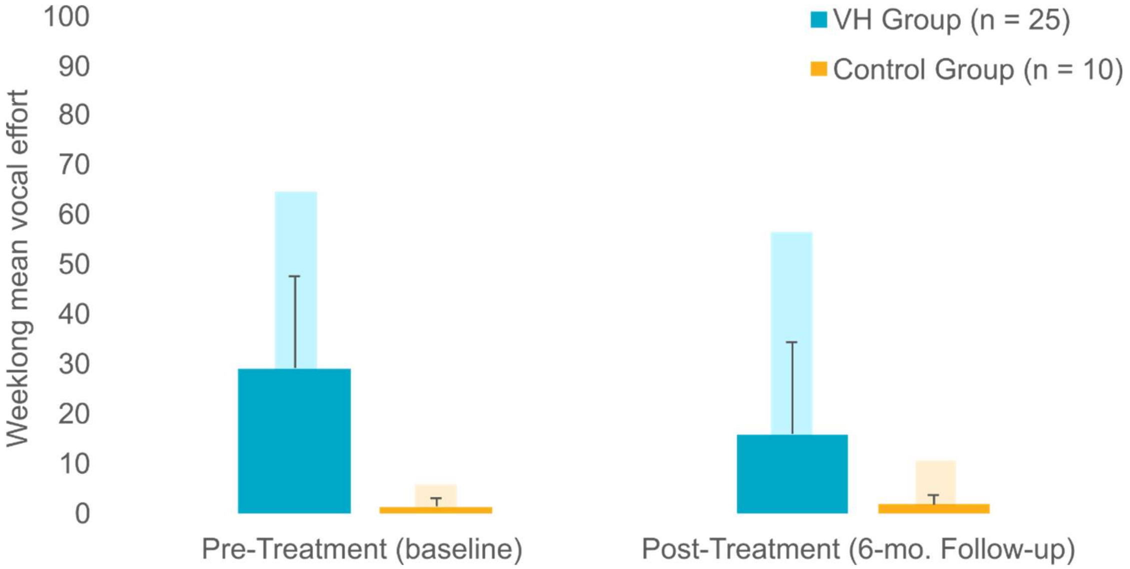
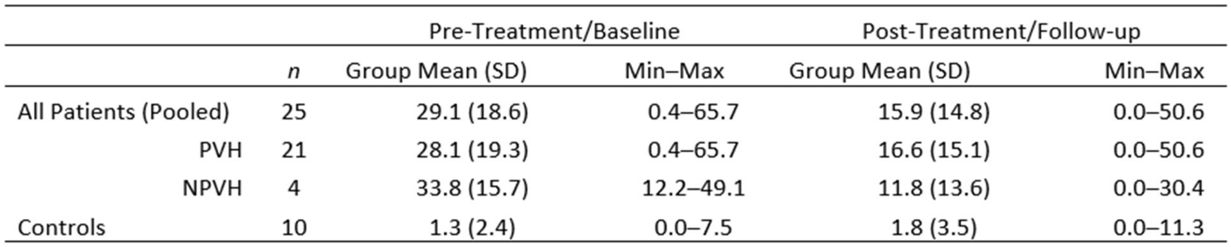


Table 5. Results for Subset 2. Group mean, standard deviation (SD), minimum (Min), and maximum (Max) of the week-long average of vocal effort scores for participants during their first week of voice monitoring (pre-treatment for patients and baseline for controls) and second week (post-treatment for patients and follow-up by at least 6 months for individuals in the control group). Results for all patients with vocal hyperfunction are shown pooled (as statistically analyzed) and by diagnosis (phonotraumatic vocal hyperfunction [PVH] or non-phonotraumatic vocal hyperfunction [NPVH]).



		Pre-Treatment/Baseline		Post-Treatment/Follow-up	
	n	Group Mean (SD)	Min–Max	Group Mean (SD)	Min–Max
All Patients (Pooled)	25	29.1 (18.6)	0.4–65.7	15.9 (14.8)	0.0–50.6
PVH	21	28.1 (19.3)	0.4–65.7	16.6 (15.1)	0.0–50.6
NPVH	4	33.8 (15.7)	12.2–49.1	11.8 (13.6)	0.0–30.4
Controls	10	1.3 (2.4)	0.0–7.5	1.8 (3.5)	0.0–11.3

Sensitivity to Change

Cohen's *d* was calculated using the pairwise comparisons of week-long mean vocal effort in individuals with and without VH, which revealed very large effect sizes (*d*= 1.62 for both individuals with PVH and NPVH compared to controls). Cohen's *d* was also calculated using data from the second analysis, comparing the scores in the patient group from pre-treatment to post-treatment for patients who were monitored before and after successful voice treatment. Cohen's *d* was 1.75, which indicated a very large effect size. The SEM (4.14) was used to obtain the MDC_{95}, which was 11.47. This finding means that for a true change to be detected, ecological vocal effort must change by around 12 scalar points.

Responsiveness

Deltas of vocal effort were calculated from each rating of participants' weeks using the global vocal status question as an index of change. The MCID was 9.30, which in this study was simply the mean absolute delta when participants indicated change in vocal status (either worsened or improved) compared to no-change scores. Table 6 lists the results for the entire dataset and also stratified by CAPE-V OS category. One outlier was removed from the WFL OS group. The stratification enabled more specificity of estimates of MDC_{95} and MCID for the WFL OS and the Mild OS groups. For the WFL group, the MDC_{95} was 5.40 and the MCID was 8.91. For the mild OS group, the MDC_{95} was 14.83 and the MCID was 9.34. There were too few participants in the moderate and severe OS groups to calculate a specific ICC (and therefore SEM), so the overall ICC of .96 was used to calculate the SEMs and therefore the MDCs listed in Table 6.

Table 6. Psychometric data for all participants, stratified by overall severity of dysphonia from the consensus auditory-perceptual evaluation of voice (CAPE-V), with score ranges in parentheses. WFL is within functional limits. Data include test-retest reliability (ICC(A,1)) of Subset 1, number of participants in each group (*n*) for all following psychometric analyses, the group standard deviation (SD) of mean week-long vocal effort during the first week of monitoring, the standard error of the measure (SEM), the minimum detectable change with 95% confidence intervals (MDC$_{95}$), and the Minimal Clinically Important Difference. ‡ indicates that test-retest reliability was not available for moderate and severe overall severities due to the small number of participants in the test-retest subset, so the overall ICC of 0.96 was used. The psychometric data for the moderate and severe groups is only meant to serve as "best-case scenario" estimates and should be interpreted with caution, as the sample sizes are very small. Psychometric data for the entire group are bolded.

Overall Severity of Dysphonia	Reliability (ICC(A,1))	*n*	Group SD of mean week-long vocal effort	SEM	MDC$_{95}$	MCID
WFL (≤10)	0.87	50	5.40	1.95	5.40	8.91
Mild (>10 and ≤35)	0.91	38	17.84	5.35	14.83	9.34
Moderate (>35 and ≤71)	‡	7	27.26	5.45	15.11	12.91
Severe (>71)	‡	5	26.37	5.27	14.61	6.54
Overall	**0.96**	**100**	**20.61**	**4.14**	**11.47**	**9.74**

Discussion

The overarching aim of this study was to examine ecological momentary assessment of vocal effort, defined as the amount of perceived work or exertion to produce voice measured in an individual's real-world speaking environment. Building on previous work that measured vocal status (Van Stan, Maffei, et al., 2017), we implemented a vocal effort prompt and VAS using the same technology via a customized platform on a smartphone device. We assessed the psychometric properties of an ecological vocal effort scale that is linked temporally to a voicing task and used to capture vocal effort ratings throughout a week of ambulatory voice monitoring in individuals with and without VH, with the goal of offering a generalizable method to measure vocal effort throughout daily life.

Test-Retest Reliability

Test-retest reliability, as it pertains to self-report of vocal status, is a challenging metric to assess; a balance must be found in spacing the ratings far enough in time so that recall effect is minimized and close enough in time to ensure that vocal status has remained constant. In individuals without VH, a time period longer than the 20-minute recording between ratings could be acceptable, as their vocal status is less likely to change; however, patients with VH may be more prone to changes in status associated with vocal demands. As noted previously, a test-retest design was not initially planned, so test-retest data was limited to the most recent year and the ratings made during in-lab procedures only. In contrast to work by Van Stan and colleagues (2017), who used Cronbach alpha to obtain an estimate of internal consistency for the construct of vocal status, we did not presume that the ecological vocal effort scale linked to a voicing task would have the same latent construct as voice-related discomfort and fatigue scales. Thus, for this study, test-retest reliability was determined to be the most appropriate measure of reliability, despite the potential limitations of the study design.

Overall, the vocal effort scores were found to be reliable based on a test-retest analysis. While it is possible that the high reliability found could be attributed to the short time-frame interval (approximately 20 minutes), a longer interval might have introduced other confounds (e.g., changes in vocal status). The test-retest reliability found using Subset 1 was .96, which indicates excellent reliability. We used a data-driven approach to identify two subgroups based on CAPE-V overall severity scores. We calculated ICCs separately for the two subgroups, finding reliability of .91 for those VH patients with mild OS and a reliability of .87 for individuals with OS within functional limits. These results suggest that the ecological vocal effort scale is reliable for individuals with OS scores under 30. Future work should determine if the reliability of the scale changes for more dysphonic individuals. Our high reliability scores of vocal effort based on test-retest reliability may indicate a “best-case scenario” in that it is the best reliability obtainable, but this means the MDC_{95} is *at least* as large as what was found, which is still valuable information. In future work, we could probe participants to re-rate vocal effort even during times when they indicate “no change” in vocal status. Although it would place more burden on the participants, this method would allow us to measure test-retest reliability in the field as opposed to only in the laboratory.

Validity

The ecological vocal effort scale was validated in the context of ambulatory voice monitoring, empirically supported by two main findings. As expected, week-long mean vocal effort was statistically different for patients with PVH and NPVH compared to individuals without VH, evidence of known-groups validity. This finding was consistent with previous results of ambulatory vocal status (Van Stan, Maffei, et al., 2017), which differentiated patients with PVH and NPVH from individuals without VH with very large effect sizes. Findings from the current study were different than those of Baldner et al. (2015), which found that the Borg CR-10 did not differentiate patients and control participants. The variability of week-long mean vocal effort scores was much larger for the patient groups compared to the control group, which indicates that vocal effort changes more throughout the day/week for patients with VH compared to individuals without VH. It is possible that patients with VH may have varying degrees of effort from one another or they may use the scale in different ways, rating in the same direction and magnitude but at different areas of the scale. Future work could investigate these patterns among patients in a more comprehensive way. Individual variability of the week-long vocal effort ratings was greater in patients compared to controls, confirming that vocal effort does in fact vary throughout the day in many individuals with VH compared to those without VH. This result corroborates other studies that have found no meaningful change after a vocally demanding event, such as fitness instruction (Dallaston & Rumbach, 2016) or singing performance (Kitch et al., 1996) in individuals without voice disorders. Our findings demonstrate the potential for tracking ecological vocal effort throughout daily life to identify instances of increased hyperfunctional behaviors. Objective measures associated with increased vocal effort could be employed as an early-warning system to either prevent VH or serve as biofeedback to aid patients in therapy as they rehabilitate, fostering carryover of strategies to natural voicing activities of daily life.

Week-long mean vocal effort was also statistically different in patients with VH after successful treatment (therapy and/or surgery), and individual variability was also reduced. Although not part of the analysis, we also confirmed that week-long mean vocal effort for individuals without VH did not statistically change from the initial week to their six-month follow-up (p = .22). These findings support construct validity for the ecological vocal effort scale and demonstrate the clinical utility of measuring ecological vocal effort throughout activities of daily living to track progress over the course of voice rehabilitation. These results were consistent with findings of reduced levels of difficulty to produce soft, high-pitched phonation, vocal fatigue, and vocal discomfort throughout daily life following successful voice treatment (Van Stan, Maffei, et al., 2017). Results from the current study corroborate findings from a study that found a statistic difference in one-time ratings of vocal effort using the Borg CR-10 before and after treatment (van Leer & van Mersbergen, 2017).

Sensitivity

The ecological vocal effort scale was found to be sensitive to the presence of vocal hyperfunction, supported by Cohen's effect sizes and the MDC_{95} (Streiner et al., 2015). Cohen's effect size is the most common measure of sensitivity in the psychometrics literature (Streiner et al., 2015). The large effect sizes found comparing individuals with and without VH, which were consistent with the effect sizes previously found for ambulatory vocal status (Van Stan, Maffei, et al., 2017). The scale was also found to be sensitive to treatment effects, also supported by the large effect size found in the validity analysis comparing week-long mean vocal effort before and after successful treatment. These findings suggest the ecological vocal effort scale has the potential to supplement assessment of a voice disorder and to document changes in vocal effort throughout voice therapy. The MDC_{95} was 11.47, which means that an ecological vocal effort score must change by at least 11.47 points for there to be true change beyond error of the measure.

Responsiveness

Whereas sensitivity evaluates a measure's ability to detect *any* change, regardless of measurement error or clinical relevance, responsiveness determines how many scalar points on the ecological vocal effort scale an individual must change for that change to be detectable beyond a margin of error. We also sought to determine the amount of change required to be *clinically* meaningful using the MCID. Our MCID was 9.30, which *should* mean that for a change in vocal status to be clinically meaningful, ecological vocal effort must change by 9.30 points; however, since the MDIC is lower than the MDC_{95} of 11.47, it is within the error of the measure and is therefore invalid. This occurrence is quite common in the rehabilitation science literature (Beninato et al., 2014; Stipancic et al., 2018), as patient-reported outcomes are challenging to measure and may not accurately reflect the measure of interest. For example, participants' ratings on the global vocal status question may not be precisely associated with ratings on the vocal effort scale, since participants were asked to rate discomfort and vocal fatigue as well. In our case, this issue is likely attributed to the intentional bias set forth, in assuming “no change” scores were deltas of zero, which prevented us from doing a receiver operating characteristic curve analysis. Other potential limitations are discussed in the limitations section. It is of some comfort that the MCID was close to the MDC_{95} even though it was still within error tolerance; instead, we must rely on the MDC as a threshold of detectable change.

To be considered a warning sign, vocal effort in patients must increase on the ecological vocal effort scale by *at least* 11.47 points; to be considered a clinical improvement, scores on the ecological vocal effort scale must decrease by *at least* 11.47 points. It is possible that a floor effect could limit our ability to detect improvements in vocal effort (i.e., decreased effort scores) in individuals who rate themselves lower on the scale. These thresholds are important as we work toward identifying objective measures of vocal function that are correlated with vocal effort and can be implemented as ambulatory voice biofeedback during natural activities of daily life. The MDC of 11.47 points (on a scale from 0–100) may suggest that an equal-appearing interval scale could be sufficient in detecting changes in vocal effort. However, more research is needed to further explore responsiveness of the ecological vocal effort scale.

When we stratified groups by overall severity (see Table 6), the WFL participants who were rated ≤ 10 in terms of CAPE-V OS (45 controls, 5 patients) had an MCID (8.91) that was larger than the MDC_{95} (5.40). This finding was not surprising, as we would expect individuals with minimal dysphonia to maintain low levels of vocal effort; an increase of 8.91 points would indeed be clinically meaningful, at least as a warning sign that hyperfunctional behaviors might be at play. Therefore, in future studies that investigate presumably vocally healthy individuals who work in at-risk occupations (e.g., teachers, telemarketers), a change threshold of 5.40 points (the MDC_{95} of the WFL group) may be enough to identify potential warning signs of VH behaviors. The metrics displayed in Table 6 for the moderate and severe groups must be interpreted with extreme caution, as the sample sizes within those groups were very small. More work is needed to understand change scores in patients with moderate and severe overall severity of dysphonia.

Limitations

As previously mentioned, a test-retest experiment was not executed from the start of the study. This limited the number of participants in the reliability analysis, as test-retest procedures were only consistently followed for the last year of the project, which happened to occur during the COVID-19 pandemic. Due to the pandemic, the hospital imposed restrictions on in-person visits, so most of the follow-up visits were held virtually. Fewer patients with VH were seen in the laryngology clinic, also due to pandemic-related restrictions. This resulted in our test-retest data including a larger number of participants who were returning at different phases of the study (post-therapy, six-month follow-up) as opposed to the first week of voice monitoring. For all follow-up weeks (for participants who had been previously monitored for a baseline week of ambulatory voice monitoring) the "in-lab" recording session was converted to a virtual format. As such, the research coordinators met with participants via a HIPAA-compliant videoconference to remind participants how to use the monitoring equipment and how to answer the vocal effort prompt and global vocal status prompts thereafter. Participants then made the voice recording using only the ambulatory monitoring equipment and a portable microphone recorder. They did however participate in test-retest the same way, rating vocal effort before and after the calibration recording. We do not believe this significantly impacted the results, as the voicing procedures and time intervals remained consistent.

An intentional bias was set forth in the in-field procedures of eliciting the vocal effort questions. Because this study was part of a much larger study that required participants to wear multiple devices and perform various system checks and calibrations each day, we wanted to reduce the hourly burden on participants as much as possible. Thus, we did not require participants to re-rate the vocal effort when they indicated that their vocal status had not changed. This inherent bias perhaps is one reason the MCID was within the margin of error of the measure and thus was invalid. Additionally, to reduce cognitive burden, we chose to display ratings from the previous hour when participants were asked to re-rate questions, so that their ratings were directly anchored to their previous ratings, which also may have imposed a bias on the data.

Future Directions and Clinical Implications

Results of this study support the use of the ecological vocal effort scale, validated in individuals with and without VH and demonstrating good reliability. The ecological vocal effort scale offers a way to measure vocal effort during activities of daily life in individuals with and without vocal hyperfunction via ambulatory voice monitoring. The intent of the authors is not to suggest that the ecological vocal effort scale is *the* best way to measure vocal effort; instead, the scale is a starting point for future work to build upon. In future related studies, the methods may be refined to reduce bias of the data, such as building in a test-retest component, potentially probing fewer times throughout the day, removing the cursors indicating previous ratings, and requiring re-ratings even when participants report no change in vocal status. The VAS was easily implemented into the smartphone platform. Following methods from the vocal status questions employed by Van Stan, Maffei, et al. (2017) no anchors, experiential or otherwise were used; future work may determine whether a different type of scale is more appropriate, such as the OMNI vocal effort scale (Shoffel-Havakuk et al., 2019), which includes pictorial and verbal anchors, the Borg CR-100, which is a logarithmic scale (Berardi, 2020; Borg & Kaijser, 2006), or a simple equal-appearing interval scale. Future work should also include participants with a greater range of severity of dysphonia or a variety of disorders beyond VH to explore a wider range of effort responses on the scale. A next step in this line of work will be to compare psychometric properties of the ecological vocal effort scale with the voice-related discomfort and fatigue scales that were also prompted for the study participants but out of the scope of the current analysis. This work may further determine whether one of these three (discomfort, fatigue, or effort) change before the others, or if all items change at the same time when a change in vocal status is detected. Furthermore, randomization of the three questions in future work could determine if an individual's answer to one question influences another.

With evidence that vocal effort changes throughout the day for patients with VH, there is benefit of tracking changes in vocal effort throughout daily life in patients with VH, with a goal of identifying objective correlates of vocal effort that could be observed during natural speaking contexts. An important next step in this line of research is to investigate vocal behaviors from the "afa" gestures, which were linked with the vocal effort ratings throughout the course of a week of ambulatory voice monitoring. Specifically, a follow-up analysis could analyze the “afa” gestures from the accelerometer signal to obtain relative fundamental frequency measures, which have been theoretically and empirically associated with vocal effort (Lien et al., 2015; McKenna et al., 2016; Stepp et al., 2010; Stepp et al., 2011). Accelerometer-derived objective measures (Espinoza et al., 2020; Lei et al., 2019; Lin et al., 2019; Marks et al., 2020; Mehta et al., 2019; Mehta et al., 2015; Švec et al., 2005; Van Stan, Mehta, Ortiz, Burns, Marks, et al., 2020; Van Stan, Mehta, Ortiz, Burns, Toles, et al., 2020; Whittico et al., 2020) may lead to better insight into the underlying physiological changes that occur when detectable changes in vocal effort are perceived. In addition to measures of vocal dose, of particular interest are objective measures that have previously been associated with vocal effort in the literature, such as subglottal pressure (Chang & Karnell, 2004; McKenna et al., 2017) and relative fundamental frequency (Lien et al., 2015; McKenna & Stepp, 2018). While the /afa/ task is not one that typically occurs in natural speech, if objective measures from the task are associated with ratings of vocal effort, future work could include the measures associated with vocal effort to capture vocal behaviors in natural running speech.

Ambulatory monitoring is changing the future of voice assessment; when systems are commercially available, voice assessment can occur outside of the clinic in real life environments, during activities of daily living. Objective correlates of ecological vocal effort could be incorporated with ambulatory voice monitoring to enable implementation of an early-warning system that could help prevent worsening of symptoms in patients with VH or those at risk of developing VH during their activities of daily living. Furthermore, ambulatory biofeedback could also be used to bring awareness of vocal behaviors associated with increased vocal effort as an adjunct to voice therapy. Biofeedback could also help patients with voice disorders carry over healthy voicing strategies learned in therapy to meet their daily vocal demands. Advances in both ambulatory voice monitoring and objective correlates of vocal effort could significantly impact the assessment and treatment of individuals with voice disorders and those at risk of developing voice disorders.

Conclusion

In the context of ambulatory voice monitoring, the ecological vocal effort scale (linked to a voicing task), was found to be reliable, valid, and sensitive to the presence of vocal hyperfunction and to successful treatment changes in vocal function in this population of individuals with vocal hyperfunction. The scale was sensitive in terms of the MDC_{95}, but not responsive in terms of the MCID. The ecological vocal effort scale offers one way to measure vocal effort in the context of daily vocal demands. Future work may determine whether the changes in vocal effort are related to vocal behaviors by investigating the objective measures that reflect the underlying physiology during times of stable or changed vocal effort, using the MDC_{95} of 12 points as the threshold for detectable change in vocal effort. For those with typical voices, a detectable change threshold was around 6 points, but would need to be closer to 9 points to be considered clinically meaningful. Additional work is needed to determine accurate change thresholds on the ecological vocal effort scale for patients with moderate and severe levels of dysphonia.

Acknowledgments

Funding provided by the Voice Health Institute and the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders (Grant P50 DC015446 awarded to REH). The paper's contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The authors thank Jarrad Van Stan, Annie Fox, Alan Jette, James Burns, and Tiffiny Hron for their contributions. We also thank Brianna Williams and Liane Houde for helping with data organization and quality checking.

References

Baldner, E. F., Doll, E., & van Mersbergen, M. R. (2015). A review of measures of vocal effort with a preliminary study on the establishment of a vocal effort measure. *Journal of Voice, 29*(5), 530-541.

Banister, E. (1979). The perception of effort: an inductive approach. *European journal of applied physiology and occupational physiology, 41*(2), 141-150.

Beckerman, H., Roebroeck, M., Lankhorst, G., Becher, J., Bezemer, P. D., & Verbeek, A. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research, 10*(7), 571-578.

Beninato, M., Fernandes, A., & Plummer, L. S. (2014). Minimal clinically important difference of the functional gait assessment in older adults. *Physical therapy, 94*(11), 1594-1603.

Beninato, M., & Portney, L. G. (2011). Applying concepts of responsiveness to patient management in neurologic physical therapy. *Journal of Neurologic Physical Therapy, 35*(2), 75-81.

Berardi, M. L. (2020). *Validation and Application of Experimental Framework for the Study of Vocal Fatigue*. Michigan State University.

Bhattacharyya, N. (2014, Oct). The prevalence of voice problems among adults in the United States. *Laryngoscope, 124*(10), 2359-2362. https://doi.org/10.1002/lary.24740 [doi]

Borg, E., & Kaijser, L. (2006). A comparison between three rating scales for perceived exertion and two different work tests. *Scandinavian journal of medicine & science in sports, 16*(1), 57-69.

Borg, G. (1982). Psychological bases of physical exertion. *Med Sci Sports Exerc, 14*(5), 377-381.

Borg, G. (1990). Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian journal of work, environment & health*, 55-58.

Bottalico, P., Ipsaro Passione, I., Astolfi, A., Carullo, A., & Hunter, E. J. (2018, 2018/03/01). Accuracy of the quantities measured by four vocal dosimeters and its uncertainty. *The Journal of the Acoustical Society of America, 143*(3), 1591-1602. https://doi.org/10.1121/1.5027816

Burke, L. E., Shiffman, S., Music, E., Styn, M. A., Kriska, A., Smailagic, A., Siewiorek, D., Ewing, L. J., Chasens, E., & French, B. (2017). Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. *Journal of medical Internet research, 19*(3), e77.

Carroll, T., Nix, J., Hunter, E., Emerich, K., Titze, I., & Abaza, M. (2006). Objective measurement of vocal fatigue in classical singers: A vocal dosimetry pilot study. *Otolaryngology--Head and Neck Surgery, 135*(4), 595-602. https://doi.org/10.1016/j.otohns.2006.06.1268

Chang, A., & Karnell, M. P. (2004, 12//). Perceived phonatory effort and phonation threshold pressure across a prolonged voice loading task: A study of vocal fatigue. *Journal of Voice, 18*(4), 454-466. https://doi.org/http://dx.doi.org/10.1016/j.jvoice.2004.01.004

Cheyne, H. A., Hanson, H. M., Genereux, R. P., Stevens, K. N., & Hillman, R. E. (2003, Dec). Development and testing of a portable vocal accumulator. *Journal of Speech, Language, and Hearing Research, 46*(6), 1457-1467. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14700368

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: erlbaum.

Cohen, S. M., Statham, M., Rosen, C. A., & Zullo, T. (2009). Development and validation of the singing voice handicap-10. *The Laryngoscope, 119*(9), 1864-1869.

Colton, R. H., Casper, J. K., & Leonard, R. J. (2006). *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Lippincott Williams & Wilkins.

Coons, S. J., Kothari, S., Monz, B. U., & Burke, L. B. (2011, Nov). The patient-reported outcome (PRO) consortium: filling measurement gaps for PRO end points to support labeling claims. *Clin Pharmacol Ther, 90*(5), 743-748. https://doi.org/10.1038/clpt.2011.203

Dallaston, K., & Rumbach, A. F. (2016). Vocal performance of group fitness instructors before and after instruction: Changes in acoustic measures and self-ratings. *Journal of voice, 30*(1), 127. e121-127. e128.

Espinoza, V. M., Mehta, D. D., Van Stan, J. H., Hillman, R. E., & Zañartu, M. (2020). Glottal Aerodynamics Estimated From Neck-Surface Vibration in Women With Phonotraumatic and Nonphonotraumatic Vocal Hyperfunction. *Journal of Speech, Language, and Hearing Research, 63*(9), 2861-2869.

Gilman, M., & Johns, M. M. (2017, 2017/01/01/). The effect of head position and/or stance on the self-perception of phonatory effort. *Journal of Voice, 31*(1), 131.e131-131.e134. https://doi.org/https://doi.org/10.1016/j.jvoice.2015.11.024

Gotaas, C., & Starr, C. D. (1993). Vocal fatigue among teachers. *Folia Phoniatrica et Logopaedica, 45*(3), 120-129.

Haley, S. M., & Fragala-Pinkham, M. A. (2006). Interpreting change scores of tests and measures used in physical therapy. *Physical therapy, 86*(5), 735-743.

Halpern, A. E., Spielman, J. L., Hunter, E. J., & Titze, I. R. (2009). The inability to produce soft voice (IPSV): A tool to detect vocal change in school-teachers. *Logopedics Phoniatrics Vocology, 34*(3), 117-127.

Hanschmann, H., Lohmann, A., & Berger, R. (2011). Comparison of Subjective Assessment of Voice Disorders and Objective Voice Measurement. *Folia Phoniatrica et Logopaedica, 63*(2), 83.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1989, Jun). Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech and Hearing Research, 32*(2), 373-392. <Go to ISI>://A1989AA47200019

Hillman, R. E., Stepp, C. E., Van Stan, J. H., Zañartu, M., & Mehta, D. D. (2020). An Updated Theoretical Framework for Vocal Hyperfunction. *American Journal of Speech-Language Pathology*, 1-7.

Hogikyan, N. D., & Sethuraman, G. (1999, Dec). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice, 13*(4), 557-569. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10622521

Hunter, E. J., Cantor-Cutiva, L. C., Van Leer, E., Van Mersbergen, M., Nanjundeswaran, C. D., Bottalico, P., Sandage, M. J., & Whitling, S. (2020). Toward a Consensus Description of Vocal Effort, Vocal Load, Vocal Loading, and Vocal Fatigue. *Journal of Speech, Language, and Hearing Research, 63*(2), 509-532.

Hunter, E. J., & Titze, I. (2008). General statistics of the NCVS self-administered vocal rating (SAVRa). *National Center for Voice and Speech. Online Technical Memo.* http://www. *ncvs. org/e-learning/tech/tech-memo-11. pdf.*

Hunter, E. J., & Titze, I. R. (2009, Jun). Quantifying vocal fatigue recovery: dynamic vocal recovery trajectories after a vocal loading exercise. *Annals of Otology, Rhinology, and Laryngology, 118*(6), 449-460.

Hunter, E. J., & Titze, I. R. (2010, August 1, 2010). Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus nonoccupational settings. *Journal of Speech, Language, and Hearing Research, 53*(4), 862-875. https://doi.org/10.1044/1092-4388(2009/09-0040)

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., & Newman, C. W. (1997). The voice handicap index (VHI) development and validation. *American Journal of Speech-Language Pathology, 6*(3), 66-70.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Controlled clinical trials, 10*(4), 407-415.

Jiang, J. J., & Titze, I. R. (1994). Measurement of vocal fold intraglottal pressure and impact stress. *Journal of Voice, 8*(2), 132-144. https://doi.org/http://dx.doi.org/10.1016/S0892-1997(05)80305-4

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009, May). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124-132. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18930908

Kitch, J. A., Oates, J., & Greenwood, K. (1996). Performance effects on the voices of 10 choral tenors: Acoustic and perceptual findings. *Journal of Voice, 10*(3), 217-227.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*(11), 1183.

Laukkanen, A.-M., Ilomäki, I., Leppänen, K., & Vilkman, E. (2008). Acoustic measures and self-reports of vocal fatigue by female teachers. *Journal of Voice, 22*(3), 283-289.

Laukkanen, A. M., & Kankare, E. (2006). Vocal loading-related changes in male teachers' voices investigated before and after a working day. *Folia Phoniatrica et Logopaedica, 58*(4), 229-239. http://www.karger.com/DOI/10.1159/000093180

Lehto, L., Laaksonen, L., Vilkman, E., & Alku, P. (2008). Changes in objective acoustic measurements and subjective voice complaints in call center customer-service advisors during one working day. *Journal of Voice, 22*(2), 164-177. http://www.sciencedirect.com/science/article/pii/S0892199706001135

Lei, Z., Fasanella, L., Martignetti, L., Li-Jessen, N. Y.-K., & Mongeau, L. (2020). Investigation of vocal fatigue using a dose-based vocal loading task. *Applied Sciences, 10*(3), 1192.

Lei, Z., Kennedy, E., Fasanella, L., Li-Jessen, N. Y.-K., & Mongeau, L. (2019). Discrimination between Modal, Breathy and Pressed Voice for Single Vowels Using Neck-Surface Vibration Signals. *Applied Sciences, 9*(7), 1505.

Liang, M. H. (2000). Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Medical care, 38*(9), II-84-II-90.

Lien, Y.-A. 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. http://dx.doi.org/10.1044/2015_JSLHR-S-14-0194

Lin, J. Z., Espinoza, V. M., Zañartu, M., Marks, K. L., & Mehta, D. D. (2019). Accelerometer-based prediction of subglottal pressure in healthy speakers producing non-modal phonation. *Proceedings of the International Conference on Advances in Quantitative Laryngology, Voice and Speech Research*.

Llico, A. F., Zañartu, M., González, A. J., Wodicka, G. R., Mehta, D. D., Van Stan, J. H., & Hillman, R. E. (2015). Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *The Journal of the Acoustical Society of America, 138*(1), EL14-EL19. https://doi.org/doi:http://dx.doi.org/10.1121/1.4922364

Marks, K. L., Lin, J. Z., Burns, J. A., Hron, T. A., Hillman, R. E., & Mehta, D. D. (2020). Estimation of Subglottal Pressure From Neck Surface Vibration in Patients With Voice Disorders. *Journal of Speech, Language, and Hearing Research*, 1-17.

McKenna, V. S., Heller Murray, E. S., Lien, Y.-A. S., & Stepp, C. E. (2016). The relationship between relative fundamental frequency and a kinematic estimate of laryngeal stiffness in healthy adults. *Journal of Speech, Language, and Hearing Research, 59*(6), 1283-1294. https://doi.org/10.1044/2016_JSLHR-S-15-0406

McKenna, V. S., Llico, A. F., Mehta, D. D., Perkell, J. S., & Stepp, C. E. (2017). Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers. *Journal of Speech, Language, and Hearing Research, 60*(12), 3404-3416. https://doi.org/10.1044/2017_JSLHR-S-17-0180

McKenna, V. S., & Stepp, C. E. (2018). The relationship between acoustical and perceptual measures of vocal effort. *The Journal of the Acoustical Society of America, 144*(3), 1643-1658. https://doi.org/10.1121/1.5055234

Mehta, D. D., Espinoza, V. M., Van Stan, J. H., Zañartu, M., & Hillman, R. E. (2019). The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *The Journal of the Acoustical Society of America, 145*(5), EL386-EL392.

Mehta, D. D., Van Stan, J. H., Zanartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., Cortes, J. P., Cheyne, H. A., 2nd, & Hillman, R. E. (2015). Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update [Original Research]. *Front Bioeng Biotechnol, 3*(155), 155. https://doi.org/10.3389/fbioe.2015.00155

Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne II, H. A., & Hillman, R. E. (2012). Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Transactions on Biomedical Engineering, 59*(11), 3090-3096. https://doi.org/10.1109/tbme.2012.2207896

Oates, J., & Winkworth, A. (2008). Current knowledge, controversies and future directions in hyperfunctional voice disorders. *Int J Speech Lang Pathol, 10*(4), 267-277. https://doi.org/10.1080/17549500802140153

Paes, S. M., & Behlau, M. (2017, Mar 9). Dosage dependent effect of high-resistance straw exercise in dysphonic and non-dysphonic women. *Codas, 29*(1), e20160048. https://doi.org/10.1590/2317-1782/20172016048 (Efeito do tempo de realização do exercício de canudo de alta resistência em mulheres disfônicas e não disfônicas.)

Popolo, P. S., Švec, J. G., & Titze, I. R. (2005, August 1, 2005). Adaptation of a Pocket PC for use as a wearable voice dosimeter. *Journal of Speech, Language, and Hearing Research, 48*(4), 780-791. https://doi.org/10.1044/1092-4388(2005/054)

Popolo, P. S., Titze, I. R., & Hunter, E. J. (2011). Towards a self-rating tool of the inability to produce soft voice based on nonlinear events: A preliminary study. *Acta Acustica United With Acustica, 97*(3), 373-381.

Qin, S., Nelson, L., McLeod, L., Eremenco, S., & Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research, 28*(4), 1029-1033.

Ramig, L. O., & Verdolini, K. (1998, Feb). Treatment efficacy: Voice disorders. *J Speech Lang Hear Res, 41*(1), S101-S116. http://www.ncbi.nlm.nih.gov/pubmed/9493749

Rosen, C. A., Lee, A. S., Osborne, J., Zullo, T., & Murry, T. (2004). Development and validation of the Voice Handicap Index-10. *The Laryngoscope, 114*(9), 1549-1556. https://doi.org/10.1097/00005537-200409000-00009

Roy, N., Merrill, R. M., Gray, S. D., & Smith, E. M. (2005, Nov). Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *Laryngoscope, 115*(11), 1988-1995. <Go to ISI>://000233839600016

Roy, N., Merrill, R. M., Thibeault, S., Parsa, R. A., Gray, S. D., & Elaine, S. (2004, Apr). Prevalence of voice disorders in teachers and the general population. *Journal of Speech, Language, and Hearing Research, 47*(2), 281-293. <Go to ISI>://000232285300004

Shewmaker, M. B., Hapner, E. R., Gilman, M., Klein, A. M., & Johns III, M. M. (2010). Analysis of voice change during cellular phone use: a blinded controlled study. *Journal of Voice, 24*(3), 308-313.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol., 4*, 1-32.

Shoffel-Havakuk, H., Marks, K. L., Morton, M., Johns III, M. M., & Hapner, E. R. (2019). Validation of the OMNI vocal effort scale in the treatment of adductor spasmodic dysphonia. *The Laryngoscope, 129*(2), 448-453.

Solomon, N. P., Helou, L. B., & Stojadinovic, A. (2011). Clinical versus laboratory ratings of voice using the CAPE-V. *Journal of Voice, 25*(1), e7-e14.

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*.

Stepp, C. E., Merchant, G. R., Heaton, J. T., & Hillman, R. E. (2011, Oct). 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*(5), 1260-1266. https://doi.org/10.1044/1092-4388(2011/10-0274) [doi]

Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2018). Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 61*(11), 2757-2771.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use*. Oxford University Press, USA.

Švec, J. G., Titze, I. R., & Popolo, P. S. (2005). Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *The Journal of the Acoustical Society of America, 117*(3), 1386-1394. http://link.aip.org/link/?JAS/117/1386/1

Tanner, K., Roy, N., Merrill, R. M., Muntz, F., Houtz, D. R., Sauder, C., Elstad, M., & Wright-Costa, J. (2010). Nebulized Isotonic Saline Versus Water Following a Laryngeal Desiccation Challenge in Classically Trained Sopranos. *Journal of Speech, Language, and Hearing Research, 53*(6), 1555-1566. https://doi.org/doi:10.1044/1092-4388(2010/09-0249)

Tenenbaum, G. E., Eklund, R. C., & Kamata, A. E. (2012). *Measurement in sport and exercise psychology*. Human Kinetics.

Tilson, J. K., Sullivan, K. J., Cen, S. Y., Rose, D. K., Koradia, C. H., Azen, S. P., Duncan, P. W., & Team, L. E. A. P. S. I. (2010). Meaningful gait speed improvement during the first 60 days poststroke: minimal clinically important difference. *Physical therapy, 90*(2), 196-208.

van Leer, E., & van Mersbergen, M. (2017, 2017/05/01/). Using the Borg CR10 physical exertion scale to measure patient-perceived vocal effort pre and post treatment. *Journal of Voice, 31*(3), 389.e319-389.e325. https://doi.org/http://dx.doi.org/10.1016/j.jvoice.2016.09.023

van Mersbergen, M., Beckham, B. H., & Hunter, E. J. (2020). Do We Need a Measure of Vocal Effort? Clinician's Report of Vocal Effort in Voice Patients. *Perspectives of the ASHA Special Interest Groups*, 1-11.

Van Stan, J. H., Gustafsson, J., Schalling, E., & Hillman, R. E. (2014). Direct comparison of three commercially available devices for voice ambulatory monitoring and biofeedback. *Perspectives on Voice and Voice Disorders, 24*(2), 80-86.

Van Stan, J. H., Maffei, M., Masson, M. L. V., Mehta, D. D., Burns, J. A., & Hillman, R. E. (2017). Self-ratings of vocal status in daily life: Reliability and validity for patients with vocal hyperfunction and a normative group. *American Journal of Speech-Language Pathology, 26*(4), 1167-1177. https://doi.org/10.1044/2017_AJSLP-17-0031

Van Stan, J. H., Mehta, D. D., & Hillman, R. E. (2015). The effect of voice ambulatory biofeedback on the daily performance and retention of a modified vocal motor behavior in participants with normal voices. *Journal of Speech, Language, and Hearing Research, 58*(3), 713-721. http://dx.doi.org/10.1044/2015_JSLHR-S-14-0159

Van Stan, J. H., Mehta, D. D., Ortiz, A. J., Burns, J. A., Marks, K. L., Toles, L. E., Stadelman-Cohen, T., Krusemark, C., Muise, J., & Hron, T. (2020). Changes in a Daily Phonotrauma Index After Laryngeal Surgery and Voice Therapy: Implications for the Role of Daily Voice Use in the Etiology and Pathophysiology of Phonotraumatic Vocal Hyperfunction. *Journal of Speech, Language, and Hearing Research, 63*(12), 3934-3944.

Van Stan, J. H., Mehta, D. D., Ortiz, A. J., Burns, J. A., Toles, L. E., Marks, K. L., Vangel, M., Hron, T., Zeitels, S., & Hillman, R. E. (2020). Differences in weeklong ambulatory vocal behavior between female patients with phonotraumatic lesions and matched controls. *Journal of Speech, Language, and Hearing Research, 63*(2), 372-384.

Van Stan, J. H., Mehta, D. D., Petit, R. J., Sternad, D., Muise, J., Burns, J. A., & Hillman, R. E. (2017). Integration of motor learning principles into real-time ambulatory voice biofeedback and example implementation via a clinical case study with vocal fold nodules. *American Journal of Speech-Language Pathology, 26*(1), 1-10. https://doi.org/10.1044/2016_AJSLP-15-0187

Van Stan, J. H., Mehta, D. D., Sternad, D., Petit, R., & Hillman, R. E. (2017). Ambulatory voice biofeedback: Relative frequency and summary feedback effects on performance and retention of reduced vocal intensity in the daily lives of participants with normal voices. *Journal of Speech, Language, and Hearing Research, 60*(4), 853-864. https://doi.org/10.1044/2016_JSLHR-S-16-0164

Van Stan, J. H., Roy, N., Awan, S., Stemple, J., & Hillman, R. E. (2015). A taxonomy of voice therapy. *American Journal of Speech-Language Pathology, 24*(2), 101-125. http://dx.doi.org/10.1044/2015_AJSLP-14-0030

Verdolini, K., & Ramig, L. O. (2001). Review: Occupational risks for voice problems. *Logopedics, Phoniatrics, Vocology, 26*(1), 37-46. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11432413

Verdolini, K., Titze, I. R., & Fennell, A. (1994). Dependence of phonatory effort on hydration level. *Journal of Speech, Language, and Hearing Research, 37*(5), 1001-1007.

Verdyuckt, I., Rungassamy, C., Remacle, M., & Dubuisson, T. (2011). Real-time embedded tracking of paitent reported vocal discomfort in professional settings. *Proceedings of the 7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*.

Vintturi, J., Alku, P., Sala, E., Sihvo, M., & Vilkman, E. (2003, Mar-Apr). Loading-related subjective symptoms during a vocal loading test with special reference to gender and some ergonomic factors. *Folia Phoniatr Logop, 55*(2), 55-69. https://doi.org/10.1159/000070088

Welham, N. V., & Maclagan, M. A. (2004). Vocal fatigue in young trained singers across a solo performance: a preliminary study. *Logopedics, phoniatrics, vocology, 29*(1), 3-12.

Whittico, T. H., Ortiz, A. J., Marks, K. L., Toles, L. E., Van Stan, J. H., Hillman, R. E., & Mehta, D. D. (2020). Ambulatory monitoring of Lombard-related vocal characteristics in vocally healthy female speakers. *The Journal of the Acoustical Society of America, 147*(6), EL552-EL558.