PREVENTATIVE VOICE MONITORING (PVM): ASSESSING THE ABILITY OF DAILY
FEEDBACK FROM DOSIMETRY TO CHANGE VOICE PRODUCTION IN
OCCUPATIONAL VOICE USERS

By

Lisa M. Kopf

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Communicative Sciences and Disorders-Doctor of Philosophy

2016

ABSTRACT

PREVENTATIVE VOICE MONITORING (PVM): ASSESSING THE ABILITY OF DAILY
FEEDBACK FROM DOSIMETRY TO CHANGE VOICE PRODUCTION IN
OCCUPATIONAL VOICE USERS

By

Lisa M. Kopf

This study explored the possibility of proactive voice dosimetry for use by occupational voice users to prevent future voice disorders. This study included two parts, the design and initial testing of a Preventative Voice Monitoring (PVM) feedback system on voice use. In Part 1, the feedback displays were designed using an iterative user-centered approach. In Part 2, testing of the design occurred in two phases: laboratory and real world (classroom). In Part 1, the researcher found that users want a layered display structure. In Part 2, the researcher found that while users felt the displays were user-friendly, users wanted an interactive, flexible system with well-defined measures that provides both objective feedback and suggestions to improve voice use. Pitch strength was found to be a statistically significant predictor of vocal fatigue, with pitch strength (associated with voice quality) decreasing with increasing vocal fatigue. Readiness to change increased from the start to the end of the study for a majority of the participants, indicating that engagement with voice monitoring and feedback increases the likelihood of engaging in behavior change. Finally, pitch and phonation time showed decreasing trends with voice monitoring and feedback, which are associated with decreased risk of developing voice disorders.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

PWP   Persons with Parkinson's Disease

PVM   Preventative Voice Monitoring

F0   Fundamental frequency

dB   Decibels

GRBAS   Grade, Roughness, Breathiness, Aesthenia, Strain

CAPE-V   Consensus and Perceptual Evaluation of Voice

CPP   Cepstral Peak Prominence

CPPS   Smoothed Cepstral Peak Prominence

ERB   Equivalent rectangular bandwidth

SOC   Stages of Change

TTM   Transtheoretical Model

PC   Precontemplation (stage of change)

C   Contemplation (stage of change)

PA   Preparation (stage of change)

A   Action (stage of change)

M   Maintenance (stage of change)

URICA   University of Rhode Island Change Assessment

RTC   Readiness to change

S-E   Self-efficacy

VFI   Vocal Fatigue Index

HRPP   Human Research Protection Program (Institutional Review Board)

Loudness	Vocal intensity (simplified name in feedback)
Quality	Pitch strength (simplified name in feedback)
Pauses	Frequency and duration of pauses (simplified name in feedback)
Strain	First moment specific loudness (simplified name in feedback)
Clarity	Cepstral Peak Prominence (simplified name in feedback)
Distance	Distance travelled by vocal folds (simplified name in feedback)
ANOVA	Analysis of variance
VLT	Vocal loading task
SPL	Sound pressure level
VBALAB	Voice BioAcoustics LABoratory

CHAPTER 1: Introduction

The Problem

One-third of adults will experience a voice disorder during their lifetime (Roy, Merrill, Gray, & Smith, 2005). Many voice disorders stem from vocal overuse and misuse (Child & Johnson, 1991), which suggests that many voice problems may be preventable. Voice disorders resulting from vocal overuse or misuse are often treated successfully through behavior modification. For example, a retrospective study by McCrory (McCrory, 2001) found that 70% of patients with vocal fold nodules experienced reduction or elimination of nodules through therapy alone. Additionally, therapists encourage clients to avoid recurrence by continuing to use the modified vocal behavior learned in therapy. Therefore, if behavior modification can treat voice disorders and prevent reoccurrence, why aren't more people changing their behavior to prevent voice disorders from occurring in the first place?

Many people may be unaware of their risk for developing a voice disorder, or unaware of what behaviors might place them at an increased risk. Approximately 25% of the U.S. workforce (over 30 million people) are at an increased risk for developing voice disorders due to their chosen occupations (Hunter & Titze, 2009; Titze, Lemke, & Montequin, 1997). This increased risk stems from the pivotal role that an individual's voice plays in his/her daily work routine. These individuals have been referred to as *occupational voice users* and include professionals such as teachers, speech-language pathologists, salespeople, actors, broadcasters, singers, receptionists, lawyers, members of the clergy, and psychologists (Hunter & Titze, 2009; Titze et al., 1997). Because these individuals' voices are an integral part of their jobs, voice disorders

would be highly detrimental to their ability to work in their chosen professions (Titze et al., 1997).

Therefore, measures should be taken to educate and facilitate behavior change for individuals at risk for voice disorders. These professionals should not have to wait until a voice problem develops before they are given support and resources for behavior modification.

A Potential Solution

Quantified self is a phenomenon where people collect data from life events to better understand and improve themselves (Choe, Lee, Lee, Pratt, & Kientz, 2014). Ubiquitous computing allows for the collection and analysis of data from personal sensors, allowing for easy access to one's own information when paired with mobile technology (Epstein, Borning, & Fogarty, 2013; Fogg, 2003).

One popular aspect of quantified self is physical activity tracking. Devices such as FitBit, Jawbone UP, and Nike+ Fuelband (Fritz, Huang, Murphy, & Zimmermann, 2014; Harrison, Berthouze, Marshall, & Bird, 2014) help individuals to track their physical activity. Many adults focus on increasing their physical activity to reduce the risk of diseases such as heart disease (Manson et al., 1999) and Type II diabetes (Hu, Sigal, Rich-Edwards, et al., 1999). Some adults simply use a device to track their physical activity over the course of the day (Fritz et al., 2014), while some use data from additional device features such as tracking calorie intake.

Physical activity trackers are a type of persuasive system. Persuasive systems are defined as: "interactive computing systems designed to change people's attitudes and behaviors" (Fogg, 2003). The persuasive nature of these systems stems from the application of principles such as self-monitoring and recognition for goal attainment (Fogg, 2003; Oinas-Kukkonen & Harjumaa,

2008). In addition, this type of system can allow individuals room to explore and experiment with strategies, similar to fish tanks and sandboxes used in video games that allow players to try different things in a relatively "safe" environment (Gee, 2008). Allowing exploration and experimentation can increase the robustness of learning, even in communication sciences and disorders (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Folkins, Brackenbury, Krause, & Haviland, 2015).

In the field of communication sciences and disorders, a number of studies have looked at clinical applications of quantified self. For example, prior research explored the use of a wearable device for monitoring saliva management in persons with Parkinson's disease (PWP) (McNaney et al., 2011). Additional studies have looked at using Google Glass for monitoring voice and speech in PWP (McNaney et al., 2014, 2015; McNaney et al., 2016).

In the area of voice science, researchers have also explored the area of quantified self. An iOS application was developed to help individuals with voice disorders monitor their home practice of voice therapy (van Leer, Pfister, & Zhou, 2016).

In addition, researchers have investigated potential applications for voice dosimeters, devices that offer long-term voice monitoring (Carroll et al., 2006; Ghassemi et al., 2014; Hunter & Titze, 2009). Like physical activity trackers, these devices can be worn daily to track vocal activity. A voice dosimeter includes an accelerometer that is worn at the neck and can either be held in place using surgical tape, or can be part of a collar that is worn around the neck (Figure 1). Dosimeters have been used in multiple contexts, including the study of voice use patterns in populations with voice disorders and high risk populations (Carroll et al., 2006; Ghassemi et al., 2014; Hunter & Titze, 2010), and for potential clinical applications (Hillman, Heaton, Masaki, Zeitels, & Cheyne, 2006; Misono, Banks, Gaillard, Goding, & Yueh, 2015). However, there is a

paucity of literature exploring the potential use of these monitoring systems as persuasive systems for voice disorder prevention.

Figure 1: VoxLog collar and Roland recorder.

The same persuasive principles that have been leveraged in preventative health using physical activity trackers could be adapted for voice use monitoring. This could reduce the need to seek voice therapy, reducing the cost to both the occupational voice user and the healthcare system. In addition, if someone does need to seek treatment for a voice disorder, the increased awareness of vocal behavior from use of this type of system may contribute to improved therapy outcomes.

The Current Study

Many individuals use self-monitoring devices, such as those that track physical activity, to allow them to observe their current performance and form future goals. This study seeks to determine whether providing feedback on voice use might have a similar impact on vocal behavior, especially for individuals who are at high risk for voice disorders. With many individuals at high risk for voice disorders, this is a potentially unmet need.

While some studies have evaluated the ability of traditional methods, such as voice hygiene education or voice therapy, for voice disorder prevention in occupational voice users with positive results (Nanjundeswaran et al., 2012; Richter, Nusseck, Spahn, & Echternach, 2016), wider implementation of these types of protocols may be restricted due to reduced motivation as well as cost. Preventative intervention by a trained voice professional may be cost prohibitive, especially if coverage is denied by insurance, which was demonstrated to be a common factor in non-adherence to voice therapy (Portone, Johns III, & Hapner, 2008). On the other hand, a voice monitoring tool may prove to be more cost effective and readily available when an individual is interested in changing their own voice.

The current study is designed to assess whether Preventative Voice Monitoring (PVM), using a dosimeter with task-based feedback, may impact vocal behavior in occupational voice users and future occupational voice users. Additionally, the results will provide preliminary evidence for which measures are most sensitive to changes resulting from PVM, including both objective measures from the voice signal and behavior change measures.

CHAPTER 2: Literature Review

Voice Use in the Workplace

When the vocal folds vibrate, they create sound. The vocal folds of occupational voice users vibrate frequently throughout the day, and many occupational voice users report voice changes associated with sustained voice use. For example, one study found that 44% of air traffic controllers, a subpopulation of occupational voice users, experienced changes in voice quality after a work shift (Villar, Korn, & Azevedo, 2016).

The changes due to prolonged voice use reported by occupational voice users can be attributed to vocal fatigue. Solomon (2008) defines vocal fatigue as "the self-report of an increased sense of effort with prolonged phonation." This increased sense of effort is likely attributed to vocal hyperfunction [voice production with "overstrained muscles" (Froeschels, 1952)] to compensate for the sensation of vocal fatigue (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1989; Solomon, 2008). However, vocal fatigue is generally transient nature, due to the voice returning to normal after voice rest (Hillman et al., 1989).

Despite the temporary duration of vocal fatigue, Hillman et al. (1989) described the hypothetical relationship between vocal fatigue, vocal hyperfunction, and the development of voice disorders. The researchers hypothesized that hyperfunction of normal vocal folds leads to vocal fatigue, with the vocal folds returning to normal with rest. However, they also hypothesized that if vocal hyperfunction persists, it can lead to the development of voice disorders. This suggests that occupational voice users, whose vocal demands often require continued voice use in the presence of vocal fatigue, are at increased risk for voice disorders.

Safety standards for occupational voice use are one way to protect occupational voice users. Safety standard recommendations have been made for exposure to multiple types of vibration in the workplace, including standards for hand and whole body vibration exposure (Griffin, 2004) and hearing protection (OSHA, 2008). However, safety standards for voice use have only recently been recommended (Titze, 2012; Titze, Švec, & Popolo, 2003).

Unlike other vibration-related standards, which are generalizable across the population, vocal dosing standards are based on multiple factors related to an individual including gender, speaking fundamental frequency, and potential genetic factors (Titze, 2012; Titze et al., 2003). Fundamental frequency (F0) is a measure of the number of cycles per second (hertz) that occur during vocal fold vibration. The perceptual correlate of F0 is pitch (Titze, 2000). The researchers determined that the interaction of these factors (gender, fundamental frequency, genetic factors) with vocal intensity and time dose determine an individual's risk of developing a voice disorder. Intensity measures the energy of a sound in decibels (dB). The perceptual correlate of intensity is loudness (Titze, 2000). Time dose measures the accumulated voicing time, the time the vocal folds are oscillating (Švec, Popolo, & Titze, 2003). This measure involves detecting each period when the vocal folds are vibrating (binary decision), and summing these periods. Time dose is related to phonation time, which is generally reported as the percent of time the person is phonating. Because of the variation of all these factors across individuals, individualized prevention measures are critical to the continued vocal health of occupational voice users.

Traditional Methods of Voice Disorder Prevention

A number of studies have explored the potential of voice disorder prevention. Many of these studies have investigated more traditional means of training, commonly referred to as

indirect and direct intervention. Indirect intervention methods involve the management of non-voice production contributors to voice disorders, and often include education on how improve vocal hygiene (Speyer, 2008). Vocal hygiene education varies between protocols, but usually involves some or all of the following elements: moderating voice use, avoiding phonotraumatic behavior such as screaming, and increasing hydration (Achey, He, & Akst, 2016). On the other hand, direct intervention involves modifying voice production (Speyer, 2008). Gartner-Schmidt, Roth, Zullo, & Rosen (2013) found that some of the most common types of direct intervention for voice therapy include resonant voice, easy voice production with vibrations felt forward in the mouth (Grillo & Verdolini, 2008), and flow phonation, voice production with appropriate airflow to allow easy vibration of the vocal folds (Titze, 2015).

In prior studies of voice disorder prevention, researchers employed indirect methods, direct methods, or a combination of the two. Indirect intervention was categorized as voice education and/or vocal hygiene education (Chan, 1994; Duffy & Hazlett, 2004; Nanjundeswaran et al., 2012). The following methods were categorized as direct interventions: vocal warm-ups (Pereira, Masson, & Carvalho, 2015), breathing training (Pereira et al., 2015), voice training (Ohlsson et al., 2015), adapted Lessac-Madsen Resonant Voice Therapy (Nanjundeswaran et al., 2012), course attendance with voice tracking (Bovo, Galceran, Petruccelli, & Hatzopoulos, 2007), and non-specific "direct training" (Duffy & Hazlett, 2004). Authors reported significant (Bovo et al., 2007; Chan, 1994; Ohlsson et al., 2015; Pereira et al., 2015) or trending (Duffy & Hazlett, 2004; Nanjundeswaran et al., 2012) improvements in voice in response to intervention. In one study reporting trends, the authors did not report statistical results due to a small number of participants (Nanjundeswaran et al., 2012).

For the other study reporting trends (Duffy & Hazlett, 2004), there were improvements on two measures (Dysphonia Severity Index, Voice Handicap Index) for individuals in the direct intervention group, and poorer performance on two measures (Vocology Screening Profile, Voice Handicap Index) for individuals in the indirect intervention group. Voice Handicap Index scores improved for the control group, with poorer scores on the other two assessments. These results suggest that direct intervention is better than indirect for voice disorder prevention. However, the reason for decreasing scores for those in indirect intervention, and improvements for those in the control group is unclear.

While there is some support for voice disorder prevention interventions, these methods often occur under the supervision of a trained voice professional. These interventions provide strategies to occupational voice users, with some interventions providing feedback on performance. However, the majority of the work is self-monitoring, and the burden is left to the occupational voice user. The outcomes from two studies (van Leer & Connor, 2012; van Leer et al., 2016) have demonstrated that technology support can increase the likelihood of home practice of therapy, and this same principle could apply to voice disorder prevention. Technology can be used as a persuasive system, a system designed to support attitude and behavior change (Oinas-Kukkonen & Harjumaa, 2008).

Persuasive Systems

The use of computers as persuasive technology, coined "captology" (Fogg, 2003), is not a new concept. Compared with human persuaders, Fogg (2003) identified multiple advantages for computers as persuaders: ubiquity (can be nearly anywhere), greater persistence in persuasion, greater anonymity for both the persuaders and those being persuaded, ability to manage large

amounts of data, and persuasion tactics can easily be scaled for a larger audience. To make the persuasive system effective, Intille (2004) suggested that messages should be easy to understand, occur at the "right" time and place, and should not be annoying or intrusive.

Fogg (2003) introduced the functional triad to categorize persuasive systems. The functional triad is used to categorize persuasive systems into three non-mutually exclusive categories: (1) a *tool* for increasing users' abilities to perform and/or analyze behaviors, (2) a *medium* for exploration and rehearsal of behaviors, and (3) a *social actor* that either engages the user directly or facilitates engagement with other users.

Many suggested principles exist for persuasive systems and the best principles to use depend on the type of system. These have been outlined in prior literature (Fogg, 2003; Oinas-Kukkonen & Harjumaa, 2008). Oinas-Kukkonen & Harjumaa (2008) summarized these principles by assigning them to one of four categories: primary task support (guiding the user to improve the target behavior), human-computer dialogue support (supporting the user in the behavior change process through rewards and suggestions), system credibility support (providing accurate and generalizable information), and optional social support (from interactions with other users). For example, tools can simplify complex tasks, tailor information for a particular user group or personalize information for a specific individual, and use self-monitoring to provide actual feedback on behavior.

However, in order for a system to be persuasive, it has to include appropriate measures for providing information and support for users. In addition, an appropriate method for obtaining the measures must be established to ensure that the measures reported are as accurate as possible. Prior work in voice science has developed tools (measures and equipment) that may be applicable in preventative voice monitoring.

Voice Dosimetry

Voice dosimeters can monitor voice use patterns. These systems typically employ, at a minimum, an accelerometer attached to the skin of the neck to capture information about the vibrations of the vocal folds. The following measures can be extracted directly from the accelerometer signal: fundamental frequency (F0), vocal intensity, and time dose (Titze et al., 2003). Increases in these measures have been established as risk factors for the development of voice disorders (Titze, 2012; Titze et al., 2003). In addition, distance dose can be calculated from these three measures. Distance dose is a measure of how far the vocal folds have traveled in centimeters, and is a mathematical combination of fundamental frequency, intensity, and time dose (Titze et al., 2003).

A number of voice dosimetry studies have observed occupational voice users over a period of days or weeks. For example, Titze, Hunter, & Švec (2007) compared periods of voicing and silence for a group of teachers over an average of 13.3 days per teacher, including both time and work and non-work time (evenings and weekends). One interesting finding was that voicing (the majority of which are less than one second) occurred an average of 1800 times per hour on weekdays, and an average of 1200 times per hour on weekends, leading to an average of 23% voicing per hour on weekdays and 13% on weekends. In another study, Schloneger & Hunter (2016) observed voice use patterns of college/university singing students over a three-day period. Data analysis included fundamental frequency, vocal intensity, and time dose, in addition to other measures. The study found significant increases in all three measures during singing compared with other times of the day.

In addition to monitoring studies, voice dosimeters are starting to be analyzed for the ability to be used in biofeedback. Van Stan, Mehta, & Hillman (2015) completed a proof-of-

concept study to determine whether vocal intensity could be used as a biofeedback measure. While the intention was to assess the feasibility of using biofeedback from dosimetry for voice therapy applications, the study was conducted using individuals without voice disorders. During biofeedback days, six participants received a vibrotactile cue (pager vibration) to reduce vocal intensity when phonation was above a certain threshold. Results of mixed ANOVA found significant reductions in vocal intensity on days with biofeedback, but vocal intensity increased to baseline levels on days without biofeedback.

Vocal dosimetry has been used primarily to study fundamental frequency, vocal intensity, and time dose in speakers. Increases in all three measures have been associated with an increased risk of voice disorders in occupational voice users. While many dosimetry studies have observed patterns in voice use in occupational voice users, newer studies support the use of voice dosimetry in biofeedback. However, while the accelerometer recordings from a voice dosimeter provide a number of measures related to risk factors for voice disorders, other measures may be able to provide information about vocal fatigue.

Need for Objective Measures of Vocal Fatigue

Currently, no objective measures have been identified that are able to reliably identify the presence of vocal fatigue (Solomon, 2008). Solomon reported that fundamental frequency is often the most correlated measure with fatigue, but for some individuals it increases with fatigue, and for others it decreases, making it difficult to determine a specific relationship. Solomon (2008) hypothesized a likely reason for the inability to reliably measure vocal fatigue is related to vocal fatigue occurring even when the sound of the voice is within the perceptually normal range and the larynx appears normal when visualized. Despite the lack of reliable objective

measures for vocal fatigue, many individuals who experience vocal fatigue report perceived changes in how the voice sounds. For this reason, it is important to understand how changes in voice quality associated with vocal fatigue may be quantified and reported to enable their use in a biofeedback application for preventative voice monitoring.

Voice Quality

In addition to research for developing safety standards for voice use, researchers are also developing objective measures that predict listener perception of voice quality. Voice disorders are often characterized as being abnormal in one or more dimensions of voice quality, including breathiness, roughness, and strain. These dimensions are subjectively rated in standard voice assessment protocols including: Grade, Roughness, Breathiness, Asthenia, Strain scale (GRBAS; (Hirano, 1981) and Consensus Auditory Perceptual Evaluation of Voice (CAPE-V; (Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009). Despite their common use in the both clinical and research settings, there is no clear consensus on the best objective measures to predict these voice quality perceptions. There is often disconnect between subjective ratings of the voice and objective measures from the voice signal. For example, patients often report changes in at least one voice quality dimension when they experience vocal fatigue (Colton & Casper, 2006; Stemple, 2000), although traditional acoustic measures have been unable to capture these self-reported changes (Solomon, 2008).

Researchers have evaluated a number of measures to assess their ability to predict listener perception of voice quality. A recent systematic review identified nine categories of measures, including both subjective and objective measures (Roy et al., 2013). The categories of objective measures were: acoustic measures (i.e., measures from acoustic recordings; Werth, Voigt,

Döllinger, Eysholdt, & Lohscheller, 2010), aerodynamic measures (i.e., measures related to airflow and air pressure; Jiang & Stern, 2004), electroglottography (i.e., measures from an electroglottogram; Baken, 1992), and image processing measures (i.e., measures from videostroboscopy and videokimography; Deliyski & Hillman, 2010).

From the acoustic signal, two types of objective measures have been explored: acoustic, measured directly from the voice signal, and auditory, measures taken after filtering the acoustic signal through an auditory processing front-end (Shrivastav, 2003; Shrivastav & Sapienza, 2003). A number of acoustic measures have been studied with both normal and dysphonic voices. Some of the more commonly cited ones are described below.

Perturbation measures capture the variability in cycle-to-cycle vibrations, and the most commonly reported are: jitter (frequency perturbation), shimmer (amplitude perturbation), and noise-to-harmonic ratio (amount of aperiodicity) (Ma & Yiu, 2005). One of the major limitations of these measures is that they are only applicable to nearly-periodic signals (Ma & Yiu, 2005). The ability of these measures to separate normal and dysphonic voices has been inconsistent. For example, Bhuta, Patrick, & Garnett (2004) looked at these three measures (in addition to other measures), and from these three, found that only noise-to-harmonic ratio predicted listener perception of quality (p= .02 for roughness, p= .007 for grade).

Newer measures have proven to be better predictors of listener perception of voice quality. One measure is the cepstral peak prominence (CPP). CPP is a measure taken from the cepstrum, the Fourier transformation of the spectrum of a voice signal (Heman-Ackah et al., 2003). A study found that smoothed CPP (CPPS) for sustained /ɑ/ vowels (CPP averaged over a number of frames) had a -0.80 correlation with grade, -0.70 with breathiness, and -0.43 for roughness. These negative correlations indicated that as each of these three dimensions

14

increased, CPPS values decreased. For running speech, the correlations were greater: -0.86 for grade, -0.71 for breathiness, and -0.50 for roughness. These findings indicate that CPPS is best at predicting overall grade, moderately good at predicting breathiness, but not well correlated with roughness perception.

Another measure is a mathematical combination of multiple acoustic measures, the Acoustic Voice Quality Index (AVQI). The AVQI is a combination of CPPS, harmonic-to-noise ratio, shimmer local, shimmer local dB, spectral slope, and tilt of the regression line through the spectrum (Maryn, De Bodt, & Roy, 2010). The authors found that this measure was able to reliably distinguish between normal and dysphonic voices ($p<.001$).

While some acoustic measures have demonstrated better predictive ability of voice quality ratings, supporters of auditory measures argued that analysis of a signal first filtered in a manner similar to the human auditory system, rather than an unfiltered acoustic signal, leads to even better matches with listener perception of voice quality (Shrivastav, 2003; Shrivastav & Sapienza, 2003).

The auditory processing front-end used in the research by Shrivastav et al. (e.g., Camacho,2007a; Shrivastav, 2003; Shrivastav, Eddins, & Anand, 2012; Shrivastav & Sapienza, 2003) was based on the model by Moore, Glasberg, & Baer (1997). This model consists of a set of filters that correspond with the series of filters in the human ear. First, an acoustic signal passes through a filter representing the outer and middle ear. This filter acts in a manner analogous to the transfer function of the outer and middle ear. The signal is then processed through a nonlinear filterbank consisting of overlapping band pass filters characteristic of the band pass filters along the length of the human cochlea, which convert the signal output from the linear frequency scale to the equivalent rectangular bandwidth (ERB) scale. An ERB is the

frequency range around a center frequency in which no change is detected by the human auditory system by small deviations in frequency from the center frequency. The ERB increases as the center frequency increases (Moore, 1983). This is in contrast with critical bands, where the critical band in which a change in frequency cannot be detected increases with increasing center frequency above 500 Hz but is constant below 500 Hz (Fastl & Zwicker, 2007). The result of the summation of the overlapping ERBs is an excitation pattern (in power units). The last step of the filter involves additional nonlinear compression, which increases the gain for frequencies above 500 Hz (Moore et al., 1997) resulting in a specific loudness pattern. The specific loudness pattern is a representation of the output from each cochlear filter as a function of frequency (in ERB units). The sum of the output from each filter represents the loudness elicited by that acoustic signal, and this is referred to as the specific loudness, which is analogous to the output to the auditory nerve.

The search for the best objective measures for signal quality is not unique to disordered voices. Fastl & Zwicker (2007) describe three primary dimensions of general sound quality: pitch strength, roughness, and sharpness. These authors define pitch strength as the saliency of pitch on a scale from weak to strong (distinct) for a given acoustic signal. This is different from pitch, which is ordered on a scale from low to high. In 2012, Shrivastav, Eddins, & Anand found a strong negative correlation (-0.989) between pitch strength and perceived breathiness in voices. These results suggest that pitch strength in general acoustic signals may be analogous to breathiness in voices.

Additionally, Fastl & Zwicker (2007) define roughness in an acoustic signal as the presence of modulation greater than 20 Hz that follows a bandpass function, the width of which is dependent on the center frequency of the tone being modulated. Along the bandpass function,

the perception of roughness increases up to a certain modulation frequency, and then decreases for increasing modulation frequencies. For example, for a 100% amplitude-modulated tone with a center frequency of 1000 Hz, the perception of roughness begins at a modulation frequency of 20 Hz and increases up to a modulation frequency of 70 Hz. From there the perception of roughness decreases until the tone is no longer heard as rough, around 300 Hz.

In addition to modulation frequency, modulation depth can influence the perception of roughness. Fastl & Zwicker (2007) demonstrated that modulation depth (extent of modulation) influenced fluctuation strength, a perceptual measure. Fluctuation strength is perceived in an acoustic signal with modulation less than 20 Hz. The relationship between fluctuation strength and modulation depth follows a sigmoid curve, with increasing modulation depth leading to an increased perception of fluctuation strength. For dysphonic voices, Eddins & Shrivastav (2013) found a similar relationship between modulation depth and perceived vocal roughness. These results suggest that the process or phenomenon that applys to the perception of roughness in most naturally occurring sounds may also apply to the perception of roughness in dysphonic voices.

Finally, Fastl & Zwicker (2007) describe sharpness as being related to an acoustic signal's envelope, and that sharpness perception is critical-band rate dependent. Critical bands are organized on a scale from 1 (lowest) to 24 (highest), where critical bands are adjoining but not overlapping, and this is referred to as the critical-band rate. Sharpness is calculated using the specific loudness of a given sound (over the critical-band range) and has additional weighting for critical-bands over 16 Bark. General acoustic signals with increased energy at higher frequencies are perceived as having increased sharpness. While sharpness is one way to capture the change in the relative distribution of energy in the spectrum, another way is through spectral moments. The

first four spectral moments are typically used to describe a spectrum, and are written with the upper case M to denote spectral moment and a subscript number to indicate which spectral moment is reported (Pinkowski, 1993). M_1 is the mean of the spectrum, M_2 is the dispersion (standard deviation), M_3 is the skewness, and M_4 is the kurtosis.

In several experiments evaluating disordered voice quality, the perception of strain has been shown to be correlated with the presence of greater energy in the higher frequencies of the vocal spectrum (Bergan, Titze, & Story, 2004; Sundberg & Gauffin, 1978). Since this change mirrors the changes observed in the perception of sharpness, one may speculate that sharpness and strain are the same (or at least, similar) perceptual constructs applied to different classes of sound stimuli.

Research in voice has also examined the relationship between strain and spectral moments. The first spectral moment measures the mean value of the energy for a given signal, and has been shown to be related to "pressedness" (strain) of the voice (Sundberg & Gauffin, 1978). Recent research found that spectral moments are better predictors of human perception when taken from an auditory rather than an acoustic signal (Kopf, Shrivastav, & Eddins, 2013). Table 1 includes all the results for each of the four spectral moments (M_1, M_2, M_3, M_4) at each stage in the auditory-processing front end, when controlling for breathiness and roughness perception. As seen in Table 1, the first spectral moment of specific loudness has been found to be the most strongly positively correlated (.832) with listener perception of vocal strain (Kopf et al., 2013).

For occupational voice users, vocal fatigue is a common report after extended voice use, and some occupational voice users report a change in the quality of their voice with fatigue. However, there are no clear "best" objective measures for capturing voice changes due to vocal

fatigue (Solomon, 2008). It is possible that auditory measures, which are more closely associated with listener perception of the voice quality, will be better predictors of voice quality changes associated with vocal fatigue.

Measure	Acoustic Signal	After First Filter	After Filter Bank (Excitation Pattern)	After Nonlinear Compression (Specific Loudness Pattern)
First Moment	0.659	0.526	0.743	0.832
Second Moment	0.303	0.611	0.013	0.309
Third Moment	0.170	0.115	0.209	-0.112
Fourth Moment	0.165	0.550	0.097	0.115

Table 1: Partial correlations between spectral moments and strain ratings.

Current Biofeedback Options for Voice

While voice dosimeters are available to capture longitudinal voice data and objective measures are being developed that are better predictors of voice quality, there is a paucity of literature looking at the application of automated biofeedback in changing voice production. One study investigated the use of real-time biofeedback in prevention of voice disorders in call center workers (Schneider-Stickler, Knell, Aichstill, & Jocher, 2012). In this study, the intervention involved PC-based software (VidiVoice) that allowed an individual to monitor vocal intensity, fundamental frequency, and phonation time by providing indications of when the voice was out of an acceptable range. The researchers report that settings were adjusted for each individual, but other than target values, few additional details were reported. For example, the target value was specified, but not an acceptable range around the target value.

The study included a treatment group receiving the biofeedback, and a control group. Biofeedback was given over the course of 4 weeks (Schneider-Stickler et al., 2012). A statistically significant decrease in Vocal Handicap Index score (lower scores indicating less

perceived vocal handicap) was found for both the treatment and control groups. The researchers also reported increased Voice Range Profile for participants in the treatment group after the intervention, as evidenced by a greater maximum dB value achieved by those who were considered to have vocal hypofunction at the start of the study. No information about participant thoughts on the software, or how much they paid attention to it during the workday were provided.

Another system that has been developed for use in the prevention of voice disorders is the Voice-Care (Astolfi, Carullo, Pavese, & Puglisi, 2015; Astolfi, Carullo, Vallan, & Pavese, 2013; Carullo, Vallan, & Astolfi, 2013). Publications have discussed the development of this small, lightweight device, compared its functioning to other developed dosimeters, and investigated its ability to monitor the voice in multiple acoustic environments. No publications regarding its functionality in voice disorder prevention have been published to date.

Behavior Change

The above information may be useful feedback to promote behavior change in Preventative Voice Monitoring (PVM). However, researchers in behavior change argue that it is not only important to account for objective changes resulting from behavior change, but the subjective, internal changes in motivation that occur as part of the behavior change process should also be accounted for (Prochaska & DiClemente, 1984). Although the observed objective changes may be subtle or difficult to determine, there may be internal changes occurring that are important catalysts in eventual observed changes (Prochaska & DiClemente, 1983; van Leer, 2010).

While there are only a few studies looking at behavior change measures in voice (e.g., Teixeira et al., 2013; van Leer, 2010), studies in other research areas have demonstrated the importance of evaluating self-change of behavior, such as in smoking cessation (Wilcox, Prochaska, Velicer, & DiClemente, 1985). In the study by Wilcox et al. (1985), the researchers were interested in the influence of participant characteristics (demographics, smoking history, health history, life experiences) in predicting self-initiated change in smoking status. The researchers found that higher scores on pleasure from smoking and smoking duration led individuals not interested in quitting to remain smokers at a 6-month follow-up. On the other hand, individuals not interested in quitting who scored high on health were more likely to quit or consider quitting at the 6-month follow-up. For a second group of individuals from the onset of the study, "relapsers," increased income increased the likelihood of quitting, and those not quitting reported smoking more daily cigarettes. Based on these findings, Wilcox et al. (1985) reported the support for their hypothesis of the importance of individual characteristics influencing the likelihood that individuals would engage in smoking cessation.

Overall, three common concepts when discussing behavior change are stages of change, readiness to change, and self-efficacy for behavior change. All three concepts help characterize where the individual is in the behavior process: their stage, how ready they are to make a change, and how confident they are that they can make changes to their behavior.

Stages of Change (SOC)

The Transtheoretical Model (TTM) is one way to describe the process of moving from lack of motivation for behavior change to taking action and maintaining changes. The TTM divides the process of behavior change into a series of five stages (Prochaska & DiClemente,

1984). The first stage is precontemplation (PC), where an individual lacks knowledge and/or motivation to make a needed behavior change. After an individual successfully moves through the precontemplation stage, they move into contemplation (C). In the contemplation stage, an individual considers behavior change but is not yet ready to initiate the behavior change process. In the third stage (preparation, PA), the individual is ready to start behavior change in the near future, but has not fully committed himself/herself to the process. When the individual starts the behavior change, he/she moves into the action (A) stage and becomes engaged in the active process of behavior change. If the individual successfully completes the action stage, they move into the final stage: maintenance (M). In the maintenance stage, the individual continues to display the changed behavior and works to avoid relapsing to previous behavior (DiClemente et al., 1991; DiClemente, Schlundt, & Gemmell, 2004). The authors added that it is possible for individuals to successfully move past the maintenance stage and keep the behavior change without active maintenance.

As an example of the stages of change, DiClemente et al. (2004) outlined the staging process for smokers. If smokers are not contemplating quitting in 6 months, they were classified in the PC stage. If they were planning to quit in the next month and had attempted to quit at least once in the last 6 months, they were assigned to the PA stage. Those that did not fit either category but were currently smoking were classified as being in the C stage. If an individual reported not being a current smoker, they were classified in the A stage if they quit less than 6 months prior and the M stage if they had quit more than 6 months prior.

In contrast, the main difference in findings between general behavior change processes and physical activity behavior change was the report of engagement in exercise at all stages of change, to differing degrees. For example, one study of SOC for physical activity (Kim, Hwang,

& Yoo, 2004) found that engagement in an exercise program leads to forward movement in stages of change for individuals in all stages below maintenance (PC, C, PA, A). In addition, a meta-analysis of 80 studies of stages of change in physical activity and exercise (Marshall & Biddle, 2001) found that even in the early stages of change, movement to later stages involved increases in physical activity. Therefore, all stages of change in increasing physical activity are marked with some increasing level of physical activity (Cardinal, 1995). This may also be true of the population at risk for voice disorders. For example, even individuals grouped in the precontemplation stage for changing vocal behavior may find themselves making some small changes in response to feedback on voice use.

A framework for the use of the TTM in voice therapy has been outlined (van Leer, Hapner, & Connor, 2008), and these authors also offered suggestions for how to move individuals in one stage of change to later stages. The authors described possible voice patients in each stage of change: PC- either not realizing that behavior contributes to their voice problem, or not interested in changing their voice; C- expressing ambivalence about current vocal behavior and behavior change; PA- interested in making changes and collaborative goal setting; A- actively engaging in behavior change inside and outside of the clinic; M- patient is independently sustaining vocal behavior changes made in therapy. The authors also described 10 behavior change processes, adapted for voice therapy, including programming reminders to practice into the patient's electronic calendar. Finally, the authors referred to some problems that may arise in the clinic related to the behavior change process and how to address them. For example, making sure that strategies taught in therapy are appropriate to the individual's stage of change. However, as stated in this article, an empirical study of the application of this framework

in voice therapy is necessary. In addition, the application of stages of change to the prevention of voice disorders was not referenced.

Five main stages of change have been identified both within a larger health-related context, and within the voice disorders population. However, the exact nature of these stages varies discipline by discipline. Whereas individuals who smoke are not interested in quitting in the PC stage, individuals in the PC stage for physical activity may engage in some light physical activity. While a framework of these stages has been developed for voice therapy, empirical support is still needed. Further work also needs to be done to extend stages of change to the at-risk occupational voice user population.

Assessing SOC

There are currently multiple methods of assessing an individual's stage of change. For example, studies such as Boyle, O'Connor, Pronk, & Tan (1998) assess stage of change through a single question with multiple answer choices, with each answer corresponding to a single stage of change. In 1983, a scale was developed to assess an individual's stage of change (McConnaughy, Prochaska, & Velicer, 1983). The researchers hypothesized five stages of change (precontemplation, contemplation, action, maintenance, and relapse), but only identified four through a principle component analysis (precontemplation, contemplation, action, and maintenance). Out of 125 items, a 32-item questionnaire was maintained, with eight items loading on each of the four factors representing the four stages of change. This scale, the University of Rhode Island Change Assessment (URICA) was developed and validated for a population of outpatient enrollees in psychotherapy (McConnaughy, Prochaska, & Velicer, 1983). The URICA has also been used in other populations such as arthritis management (Keefe

et al., 2000) and weight loss (Prochaska, Norcross, Fowler, Follick, & Abrams, 1992). Finally, the URICA has been adapted for use with individuals with voice disorders (URICA-VOICE; (Teixeira et al., 2013), and was used as an outcome measure in a study of teachers with voice complaints (Rossi-Barbosa, Gama, & Caldeira, 2015).

While there are multiple ways to evaluate stages of change, one (URICA-VOICE) has been developed specifically for use with the voice disorders population. However, its application for the at-risk occupational voice user population still needs to be investigated.

Readiness to Change (RTC)

While the URICA-VOICE can assess stages of change, it can also be used to assess individuals' readiness to change. Readiness to change (RTC) is described as the interaction of how important an individual thinks a problem is and how confident he/she feels that a change can be made (DiClemente et al., 2004). Even though it is related to stages of change, this concept is a separate entity. RTC has been used as a predictor of future behavior change: the greater the RTC, the more likely the individual will engage in behavior change (DiClemente et al., 2004).

Readiness to change can be calculated arithmetically from the results of the URICA (DiClemente et al., 2004) and the URICA-VOICE (Teixeira et al., 2013). Readiness to change was discussed in the paper by Teixeira et al. (Teixeira et al., 2013), with individuals in later stages of change demonstrating greater readiness to change. Another study used the readiness to change measure from the URICA-VOICE and concluded that a majority of teachers with vocal complaints have low readiness to change and are in the pre-contemplation stage of change (Rossi-Barbosa et al., 2015).

Readiness to change is related to SOC, but is a different concept. RTC has been explored in the voice disorders population, but further exploration is needed, especially in the at-risk occupational voice user population.

Self-Efficacy (S-E)

Related to readiness to change is self-efficacy. Self-efficacy for behavior change is a measure of an individual's confidence in his/her own ability to successfully change his/her own behavior (Bandura, 1977). As individuals transition to later stages of change, self-efficacy for behavior change also tends to increase (DiClemente et al., 2004; Dijkstra, De Vries, & Bakker, 1996). Researchers found that larger increases in self-efficacy occur moving from preparation to action and from action to maintenance than forward progression in the earlier stages of change. The researchers attributed these larger increases in self-efficacy for later stages to individuals' successful engagement in behavior change, whereas those in earlier stages were not yet ready to engage in behavior change.

Self-efficacy has been described as an important construct to evaluate in individuals with voice disorders (van Leer, 2010; van Leer & Connor, 2010, 2012). A scale was designed to assess S-E for voice therapy (van Leer & Connor, 2010, 2012), but this scale focuses on barriers to voice therapy home practice, and does not cover the more broad topic of voice change that may occur outside of voice therapy, such as through preventative measures initiated by the individual himself or herself.

While S-E has been evaluated for multiple types of behavior change, there is a paucity of literature exploring it in the voice disorders population, and its application to the at-risk occupational voice user population remains an area of needed study.

Vocal Fatigue Index (VFI)

Stated previously, there are a number of factors to consider with vocal fatigue, including subjective and objective measures. One questionnaire has been developed to assess vocal fatigue, the Vocal Fatigue Index (VFI). As discovered by Nanjundeswaran, Jacobson, Gartner-Schmidt, & Verdolini Abbott (2015) in the development of the Vocal Fatigue Index, vocal fatigue appears to be divided into three factors. Factor 1 is characterized by a feeling that the voice is "tired," which may lead to a reduction in further voice use. Factor 2 is characterized by physical discomfort, such as a sore throat. Finally, Factor 3 is characterized alleviation of fatigue symptoms with vocal rest. These findings are not surprising given the complex nature of vocal fatigue, and highlight the need to identify the most important factors for assessing the voices of occupational voice users.

CHAPTER 3: Study Aims & Hypotheses

The goal of the current study was to determine whether Preventative Voice Monitoring (PVM), using a dosimeter with task-based feedback, impacts vocal behavior in occupational voice users and future occupational voice users. Three study aims were defined: to design the feedback (Aim 1), and to assess its ability to influence both behavior change measures (Aim 2) and voice production (Aim 3). To address the aims, the study was divided into two parts: the creation of the feedback (Part 1) and testing of the feedback (Part 2). The experimental methods, data, and results are explained in greater detail in Chapters 4 and 5. This was done to explain the temporal sequence in which the study was conducted.

Aim 1: To extract design requirements for conveying feedback to users.

The formative evaluation was exploratory in nature. In Part 1 of the dissertation, a qualitative, interpretivist approach was used to identify the most relevant objective measure(s) and type(s) of visual displays for potential users. The researcher used iterative prototyping (Bernstein & Yuhas, 2005; Goldman & Narayanaswamy, 1992) to design the best delivery method for the feedback. For each iteration, the information gathered from additional participants shaped display prototypes for the next iteration. One participant returned to check the final iterations of the feedback to determine their appropriateness. In Part 2, participants used the feedback displays with vocal loading tasks. Interviews provided insight into future iterations of the feedback displays. Interviews after baseline recordings (no feedback) provided additional information on occupational voice user needs. Final interviews provided insight into aspects of the feedback that worked well and those that need further development.

Aim 2: To identify changes in voice behavior management after receiving feedback.

Increased RTC and S-E are linked to increased engagement in the behavior change process (DiClemente et al., 2004; Dijkstra et al., 1996; Marshall & Biddle, 2001). In addition, voice training can reduce self-perceived vocal fatigue symptoms (McCabe & Titze, 2002). In Part 2, RTC, S-E, and vocal fatigue were assessed at three time points: at the beginning of the study (prior to vocal loading tasks), after baseline recordings (with no feedback), and at the completion of the study (after feedback recordings). It was hypothesized that active engagement in changing voice production would manifest as improvements in: RTC (increase), S-E (increase), and vocal fatigue. The improvements in vocal fatigue were defined as changes in one or more of the following of the VFI (consistent with Nanjundeswaran et al., 2015): a decrease in Factor 1, a decrease in Factor 2, and/or an increase in Factor 3. In addition to these assessments, interview results after baseline recordings provided further insight to what behavior changes were made intuitively, and final interviews provided insight into what changes were facilitated by the current feedback.

Aim 3: To quantify changes in the voice after receiving feedback.

This aim consists of three hypotheses:

1. Vocal intensity, time dose, and fundamental frequency (F0) have been identified as factors contributing to the distance travelled by the vocal folds during speech (Švec, Popolo, & Titze, 2003; Titze et al., 2003). These prior studies reported that decreasing the distance travelled by the vocal folds (distance dose) reduces the likelihood of damage to the vocal folds. In Part 2 of the current study, participants completed vocal loading tasks

and received feedback on voice production using the feedback displays developed in Part 1. It was hypothesized that occupational voice users would improve voice production in response to feedback, and these improvements would manifest as decreases in one or more of the following: vocal intensity, voicing time, and/or F0.

2. When patients report vocal fatigue, they often describe perceived changes in voice quality, including increased breathiness and vocal effort (Colton & Casper, 2006; Stemple, 2000). Part 2 participants produced three sustained /ɑ/ vowels before and after each vocal loading task, and average values of pitch strength and first moment specific loudness were compared. It was hypothesized that increasing vocal fatigue would result in the following changes in objective voice quality measures: increasing strain (increasing first moment specific loudness) and/or increasing breathiness (decreasing pitch strength).

3. It was further hypothesized that the changes in breathiness and vocal strain pre- to post-vocal loading task would be greater for baseline tasks than tasks with feedback. This was anticipated because vocal fatigue should decrease with behavior change related to feedback, and therefore should lead to smaller increases in breathiness and/or vocal strain.

CHAPTER 4: Feedback Requirements and Changes in Voice Behavior Management (Aims 1 & 2)

Study Overview

The goal of this study was to determine whether Preventative Voice Monitoring (PVM) would impact vocal behavior in occupational voice users and future occupational voice users. Additionally, the results would provide detailed user insight into the design and implementation of PVM (Aim 1), and insight on the sensitivity of behavior change questionnaires to internal changes experienced by users of PVM (Aim 2). This user-centered design study, a study involving the end user throughout the design and testing process (Beyer & Holtzblatt, 1998; Lucas Jr., 1971; Slegers, Duysburgh, & Jacobs, 2010), consisted of two parts (Table 2).

Study Part	Iteration/Phase	# Sessions (per participant)	# Unique Subjects
Part 1	**Feedback Display Prototype Development**	**1**	**15**
	Initial Feedback Display Prototypes		6
	Iteration 1 of Feedback Display Prototypes		3
	Iteration 2 of Feedback Display Prototypes		3
	Iteration 3 of Feedback Display Prototypes		3
Part 2	**Feedback Display Prototype Testing**	**11**	**18**
	1: Laboratory Testing		13
	2: Field Testing		5

Table 2: Study design overview.

The length of each part of the study and number of participants enrolled were appropriate based on literature in both communication sciences and disorders and human computer interaction. In a systematic review of voice therapy literature on the effects of therapy, 14 of the

47 included studies had 20 or fewer participants (Speyer, 2008). In a study of an application for voice monitoring, 14 individuals enrolled in voice therapy piloted an application and completed short interviews to assess their perceptions of the application (van Leer et al., 2016). A study of a Google Glass application to monitor vocal intensity for persons with Parkinson's disease (PWP) by McNaney et al. (2015) followed a similar structure to the current study. This study consisted of three phases: (1) a design workshop involving 7 PWP, (2) a 30-minute pilot of the application with 8 individuals who did not have Parkinson's disease, and (3) a three-day field trial of the application with 6 PWP. Consolvo et al. (2008) recruited 12 individuals to complete a study of a physical activity tracker. This study involved individuals coming in for three interviews (initial, midpoint, final), and using the physical activity tracker for a total of three weeks. Another research group tested a physical activity tracker with 8 participants over a month, one week of baseline measurements and three weeks with feedback (Lim, Shick, Harrison, & Hudson, 2011).

Part 2 of the current study was similar in length to the Lim et al. (2011) study. Participants in Part 2, Phase 1 (Part 2.1) were scheduled three days a week (Monday, Wednesday, Friday) at the same time each day, and the estimated length of time to complete the study, with no missed sessions, was four weeks. The researcher chose this scheduling format because it mirrored teaching schedules for instructors who taught a class three times per week. Due to sickness and other excused absences, all participants completed Part 2.1 within six weeks. Participants in Part 2, Phase 2 (Part 2.2) were scheduled on teaching days for lecture-based courses. Participants were included in the study if they taught at least two different days per week. All Part 2.2 participants taught two days a week (Monday, Wednesday), and the estimated time to complete the study was six weeks. Due to sickness and other excused absences, all participants completed Part 2.2 within ten weeks. One participant in Part 2.2 (P2205) only

completed seven of the eight recording sessions due to a change in course format partway through the semester. Due to this change, only one day per week was recorded for the final few weeks because the other day was primarily student presentations and little instructor talking.

Inclusion Criteria

All participants were recruited from the Michigan State University community through posted flyers and word of mouth. The same inclusion criteria were used for all participants in the study and are included in Table 3. While being an occupational voice user was a criterion for inclusion, this particular criterion was not limited to previously established occupational voice user categories, but was broadened to include any occupation where the participant felt that talking was an integral part of the job requirements.

Category	Inclusion Criteria
Occupation	Occupational voice user or student studying to be occupational voice user
Age	18-65 years old
Hearing Screen	Pass at 0.125 to 8.0 kHz at 20 dB (ANSI S3.21-2004)
History	No history of (or current) voice disorders requiring medical intervention

Table 3: List of inclusion criteria.

Part 1 Methods

Part 1 involved formative evaluation of design requirements for feedback using iterative paper prototyping. In the iterative design process, the researcher presented iterations (versions) of display images to potential target users (people that the product is being designed for), and asked potential target users' opinions and suggestions for how to improve the displays (Bailey, 1993; Buley, 2013; Hansen, 1997). The researcher used hand-drawn sketches in prototyping because sketches welcome more conceptual commentary than more polished computer-generated

images, which may evoke more commentary on formatting (Buley, 2013). Changes in later iterations were based on input from previous potential target users. Part 1 consisted of four iterations, and displays were finalized at the end of the fourth iteration. An overview of the organization of Part 1 can be seen in Figure 2.
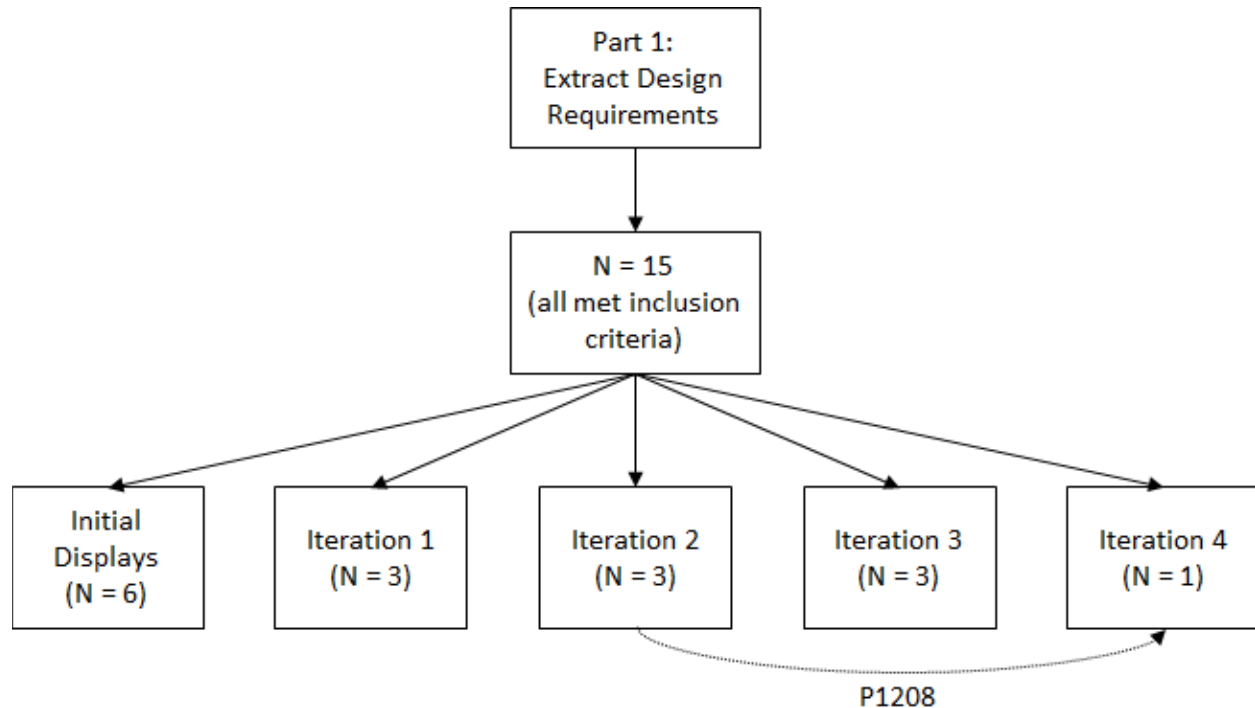
Figure 2: Outline of Part 1.

Participants

Fifteen participants enrolled in Part 1 (Table 4), (4M, 11F; M=42.0 years, SD=15.38 years, Range=22-64 years). Two of the older participants (P1209, P1414) failed the hearing screen at one or more of the higher frequencies (over 4000 Hz) in one or both ears. However, their hearing did not appear to affect their understanding of conversational speech, and therefore, the researcher included these participants in the study. The researcher assigned participants to

one of the four iterations based on the date of their participation, and the iteration is reflected in the participant number (e.g., participant numbers starting with 11 viewed the initial displays, participant numbers starting with 12 viewed the second set of displays, and so on).

Participant ID	Age	Gender	Occupation
P1101	35	F	Course Instructor
P1102	27	F	Office Assistant
P1103	30	F	Office Assistant
P1104	35	M	Podcast Producer
P1105	32	F	Teaching Assistant, Student
P1106	54	F	Lawyer
P1207	64	F	Course Instructor
P1208	27	F	Teaching Assistant, Student
P1209	56	F	Library Clerk
P1310	22	M	Office Assistant, Student
P1311	56	F	Research Assistant, Student
P1312	62	F	Radio Volunteer
P1413	26	M	Course Instructor
P1414	63	M	Retail, Substitute Teacher, Santa Claus
P1415	41	F	Research Assistant, Student

Table 4: Part 1 Participant Demographics.

Procedures

Each participant enrolled in an individual session of no more than 1.5 hours in duration. Prior to participation, all participants first completed an informed consent form approved by the Michigan State University Human Research Protection Program (HRPP). Then, each participant completed an intake form (Appendix A). After inclusion criteria were met and the initial forms were completed, the remainder of the session was recorded using two digital recorders to ensure no data loss (Roland R-05, Lake Stevens, WA, USA; TASCAM DR-40, Montebello, CA, USA). In the semi-structured interview, the researcher had a series of printed questions, but asked follow-up questions when relevant to gain a deeper understanding of a participant's answers. For

example, if a participant reported that their environment was only occasionally noisy, the

researcher might ask her/him to describe situations when this occurs.

   After the initial interview (Appendix B), the researcher showed the participant a set of

icons (Figure 3), one representing each of the objective measures to be tested. Icons were

randomized between participants to ensure that presentation order did not have an effect. The

researcher assigned simplified names to these objective measures to ensure easier understanding

by participants. These measures were: vocal intensity (loudness), frequency/duration of pauses

during speech (pauses), distance travelled by the vocal folds (distance), pitch strength (quality),

and first moment specific loudness (strain). A sixth measure, cepstral peak prominence (clarity),

was only used with the first participant because while this measure is different from quality, the

basic explanations of the two measures were so similar that it caused confusion. Explanations of

measures were provided as needed. The researcher first provided a simple definition of the

measure, and if prompted by the participant, expanded on the definition as needed.
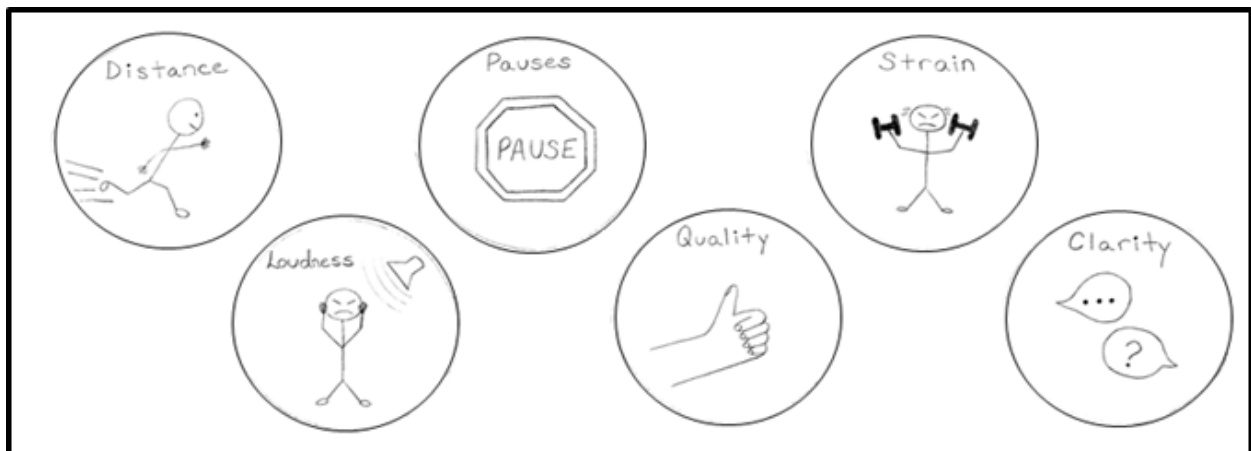
Figure 3: Initial icons representing objective voice measures. Icons and the number of measures
changed in later iterations based on participant input.

After the introduction of the measures, the researcher instructed the participant to order the icons from most useful to least useful, and assign each a number (1=not useful, 10=the most useful). The participant could assign multiple measures the same usefulness. The researcher then encouraged the participant to describe why she/he assigned the given order to the measures. Participants also indicated how important they thought each measure was, and how confident they were that they could change a given measure based on daily feedback using the same scale.

Next, the researcher encouraged participants to provide initial insight on the design of the feedback displays without seeing the initial prototypes. The researcher provided options for responses: verbal, drawn using a pencil and paper, or a combination of the two.

Finally, the researcher introduced prototypes of visual feedback displays for measures of interest one measure at a time in a random order. The participant saw the displays sequentially and then simultaneously. During this process, the researcher asked a series of questions about the visual feedback displays (Appendix C). Questions for displays included asking participants what kind of information they are able to gather from a display (without researcher explanation), whether they can identify general patterns in the display, whether the display itself is useful, and if the display should change (and how). Additional questions included whether the participant would combine displays, and whether the participant could offer any additional design ideas. After displays of individual measures, participants saw one or more examples of multi-measure displays.

Part 1 Analysis

Aim 1

To extract design requirements for conveying feedback to users.

In Part 1, the researcher conducted qualitative analysis of participant comments and suggestions following the six individual sessions with the initial feedback displays. This analysis allowed the researcher a better understanding of the needs of occupational voice users, and provided guidance in determining the most relevant measures, feedback displays, and display edits to incorporate into updated prototypes for the next iteration. The most common suggestions were incorporated, as well as any additional insightful suggestions judged worthy of further evaluation. The same analysis and iterative design was completed for each successive iteration. After determining the finalized displays, the researcher brought one participant from an earlier iteration back to look over the displays. This participant was particularly insightful in the earlier session, making her a trusted individual for a final check of the displays.

Part 1 Results

The results from Part 1 are reported below. The results highlight the major changes in feedback displays from one iteration to the next based on participant input.

Initial Feedback Designs

The initial feedback display designs can be seen in Appendix D. These initial designs were based on recommendations from the literature on data visualization (Holmes, 1984; Tufte, 1983, 2001). For example, the shape used for the daily feedback is an example of a golden

rectangle (Tufte, 1983, 2001), a rectangle with dimensions that have been shown to be highly pleasing to the eye. In addition, limited "data ink" (Tufte, 1983, 2001) was used in these designs, as evidenced by the white bars in the bar graphs with only a black outline. In addition, the color palate was limited to black ink on white paper, except for occasional red to mark a particular contrast (Tufte, 1983, 2001). Finally, some of the designs, such as the speedometer for distance, were designed to look like a common object measuring a similar phenomenon (Holmes, 1984; Tufte, 1983, 2001).

Iteration 1

Two measures were omitted (distance, clarity). Clarity was merged with quality after the first participant (displays were all presented as quality to the subsequent 5 participants) because the explanation of these measures was extremely similar, even though they are unique objective measures. Distance was removed because, while participants occasionally asked for clarification about one or more measures during the study, each of the first six participants asked for clarification on the meaning of distance between 1-3 times during the interview. Therefore, the researcher determined that this measure may be too abstract for simple voice feedback.

Other changes included introducing dynamic icons. Many participants commented that icons could be a way to track performance, and so the researcher presented two versions of each icon in Iteration 2. In addition, the pause icon will feature the pause count in the center of the stop sign. Participants also indicated that the loudness icon was confusing- the sound should be coming from the stick figure, not the wall, to convey that the system is measuring the speaker's loudness. Participants felt strain should include with two features: lightning bolts from neck and position of dumbbells (lower for less strain).

Icon Suggestions

P1102 introduced the idea of dynamic icons through describing appropriate changes for the icons (Appendix E, Figures 61-64). For example, she suggested the face would change for strain to demonstrate "more relaxed or more effort" (Figure 64). In addition, pauses be larger for more pauses, and smaller for less pauses.

Changes for all icons were then shaped by the idea of having dynamic icons. P1101 suggested, "Loudness icon doesn't convey that you're being too loud. Loudness should be person talking with audience holding their ears." However, this idea would involve a greater amount of data ink, making the icon more complex. Therefore, loudness was changed to look more like other devices, such as more lines or dots on a radio or cell phone. However, the lines were drawn to be coming out of a person's mouth, with a wider mouth opening for louder speech (Figure 61).

For pauses, most participants pointed out that pauses included count and length of time. While P1102 had suggested changing the size of the sign to indicate pauses, this might be harder for individuals using the device to tell how many pauses. She did add, "I like the icon because pause looks the same as stop." P1106 added, "This would be a count- how many times you paused." So, the stop sign was kept, but it was simplified to replacing the word "pause" to the number of pauses made (Figure 62).

The idea for the change in thumbs in the quality icon was introduced by P1105: "You are going to use thumb up/down or just thumb up? One image is not enough- this image indicates good voice. It's good, but need to add more levels." Therefore, the dynamic icon included varying the thumbs up and/or down (Figure 63).

Finally, there was feedback on the strain icon. P1105 wanted the icon to better reflect what it was measuring, "Maybe use a picture of a person who is talking, lips moving." She also

added: "Showing someone using a lot of effort in talking." Therefore, strain was changed to having an open mouth with lines coming from the neck to indicate increasing strain (Figure 64).

Loudness Suggestions

Overall, participants preferred the sparkline: a small line representing a lot of information (Tufte, 2001) which can be seen in the bottom display for loudness in Appendix D (Figure 47) and top display for loudness in Appendix E (Figure 65). This was displayed as a small multiple-multiple, with time periods presented in miniature so all can fit (Tufte, 1983, 2001). This display was preferred over the bar graph (Figure 45) and the clock (Figure 46). P1102 commented that the sparkline is the best "if the objective is to track and adjust," and P1104 remarked, "This makes more sense to me [than another loudness display]." P1106 liked the use of the color red in the displays because "Red usually gets people's attention that something is wrong."

However, in the next iteration, the small multiple display was expanded to allow easier view of daily detail (Appendix E, Figure 66). This change was based on comments such as "I like the 3rd one the way it is, maybe a little bigger" from P1101.

Despite liking the sparkline, participants felt that this display held a lot of information, so it might also work as a display for individual day rather than a small multiple. Therefore, this idea was added as a second display in the next iteration. As P1105 put it, "Even if you have one day's voice tracked in one figure, that's a lot of information."

The clock and stacked bar graph were not liked overall, so they were omitted from the next iteration. For example, P1102 found the clock the most difficult to understand, and P1103 commented, "I don't know what's happening here." P1104 said that the clock "display is nice, but information is difficult to comprehend." P1103 did not like the bar graph: "It takes way too

41

much work to understand what's happening," and P1104 added, "I really don't understand the red or the gray [in the bar graph]."

Additionally, a "pop-up" feature was included to give users a definition of measures (presented on a sticky note, not included in figures). This was based on feedback from P1102: "But I would like to be able to go back and check definitions/hit the icon for definition."

Pause Suggestions

Participants preferred the small multiple line graph (Appendix D, Figure 48), although some liked the clock as well (Appendix D, Figure 49). Both were preferred over the stacked bar graph (Appendix D, Figure 50). Therefore, a display was created that incorporated the two (Appendix E, Figure 68), and a second display was created without the clocks (Appendix E, Figure 69). P1104 liked the small multiple (Appendix D, Figure 48) and commented, "Should be easy to track over time because we work in comparisons." P1101 commented that the clocks could be "Helpful to see overall how much break time you give students." P1105 felt that both the clock and the small multiple would be helpful, with the clock as a real-time measure and the small multiple as a summary measure across multiple days. P1104 was one of the participants who did not like the stacked bar graphs because: "Having 2 sets of bars is difficult."

Another display (Appendix E, Figure 67) was created that involved tick marks for each pause, with each tick of varying length to represent pause duration. This was based on a comment from P1106 that she liked the small multiple but wanted hash marks to indicate when pauses occurred.

Quality Suggestions

There were six displays tested for quality because there were initially three designs each for quality and clarity. Out of the quality/clarity displays, participants preferred the smiley faces (Appendix D, Figure 51; Appendix E, Figure 70). P1102 commented that the smileys were "very self-explanatory," and compared with the other quality displays, P1103 stated: "I liked the smileys the best." P1104 added that the smiley display was best because "people like to spend less time reading."

Based on feedback in favor of the clarity line graph (Appendix D, Figure 54) and the quality line graph (Appendix D, Figure 53), another display was created (Appendix E, Figure 71). P1105 felt that the smileys and the line graph (Appendix D, Figure 54) were the two best displays for quality because they give different information and the line graph giving "more information than the smiles." On the other hand, P1106 found the smileys "hokey," and preferred the line graph for quality. The researcher added color (red) to determine whether this enhanced the display based on the early comment from P1106 about loudness. P1101 suggested displaying combining smileys and numbers on axes, which was incorporated into the design (Appendix E, Figure 71).

Finally, a new version of the small multiple quality display (Appendix D, Figure 52) was created with greater time detail (Appendix E, Figure 72). This design was kept based on a comments from P1106, "One thing I like about this one is that it shows your change each day."

Strain Suggestions

P1101 liked the line graphs for strain (Appendix D, Figure 58), but suggested the researcher "add numbers to label endpoints" (Appendix E, Figure 74). P1104 suggested varying

the color along the lines. Therefore, the researcher tried to differentiation by varying the lines'

color on the gray scale (Appendix E, Figure 73).

Multi-Measure Suggestions

Finally, the researcher simplified the multi-measure display to include the remaining four

measures (Appendix E, Figure 75). While some liked the multi-measure display, P1103

commented, "I would rather look at one [measure] at a time on a daily basis." P1105 commented

that the multi-measure display was "too big for a mobile phone." Despite mixed reviews for a

multi-measure display, it was kept in future iterations because opinions were split.

Iteration 2

As a result of feedback on the first iteration of displays, incremental changes were made

to all feedback displays.

Icon Suggestions

P1207 was not a fan of the dynamic icons, stating, "Don't change background icons, just

thumbs up/thumbs down." P1209 was also disinterested in dynamic icons: "Icons are cute, but

looking for overall, I won't want to work so hard. I'd rather have the graph and be able to zoom

in." P1208 liked the strain icon, "I think lifting weight is good. Lightning from neck is cute. Man

changing colors would be good: hotter colors for more strain." Comments like this from P1208

indicated that dynamic icons might be worth a further look, so they were maintained in the

second iteration (Appendix F, Figures 76-79).

Loudness Suggestions

Based on feedback, the small multiple was kept but not the elongated one (Appendix F, Figure 81). As P1209 stated, the elongated multiple was "Intimidating if looking for general pattern," and P1207 said it "Seems scary busy". P1208 agreed: "I like short format better than long." P1209 liked the idea of the pop-out, but wanted it to be better labelled than simply clicking on one of the images on the display: "I wouldn't think to click on the axis to see these explanations." Therefore, a small question mark icon was added.

While the elongated multiple was not preferred, P1209 remarked, "I would want to be able to zoom in like this." In other words, the participant wanted to be able to zoom into a single elongated day to be able to see additional detail (Appendix F, Figure 80).

Pause Suggestions

Participants liked the idea of the tick mark display as a pop-out (Appendix E, Figure 67), so it was retained (Appendix F, Figure 83). P1207 commented, "You don't want this to be the main screen because too busy," but that it was good as a pop-out.

While P1207 and P1209 were not very interested in the clocks, P1208 was excited about them: "It's a clock! I like this a lot. I can see making a game out of the clocks [extending pause time]." Therefore, the clocks were retained (Appendix F, Figure 84). P1209 was not interested in the line graph, and felt that "seeing it as a number would work." This update was made to the clock display.

The line graph was retained (Appendix F, Figure 82), although simplified, because P1209 commented, "The line makes more sense to me," which is a sentiment shared by prior participants as well.

45

Quality Suggestions

Out of the quality/clarity displays, participants P1207 and P1208 preferred the smiley faces (Appendix F, Figure 85). P1209 commented, "Smileys are too simplistic." P1209 voted to keep the line graph. While not liking the smileys in their own display, P1209 remarked, "I like incorporating emojies & numbers [on the axis]." Therefore, the single line graph was kept but reduced to a single day to act like a pop-up (Appendix F, Figure 87), and used in a small multiple display as well (Appendix F, Figure 86).

Strain Suggestions

P1208 and P1209 did not like the arrows at the endpoints. P1208 liked the numbers of the endpoints better, and P1207 remarked, "Redundant numbers on endpoints can be useful to non-scientists." Therefore, the endpoint numbers were retained, and not the arrows (Appendix F, Figure 88).

The second design (Appendix I, Figure 89) was based on a suggestion from P1207, "Parallel lines. One line is for morning, one is for evening." This comment led the researcher to recall a comment from an earlier participant for labelling the lines, P1101, "Use morning/night as markers on axis."

Multi-Measure Suggestions

No changes were made to the multi-measure display (Appendix F, Figure 90). Again, there were mixed comments about retaining it. P1207 preferred the idea of an iconic dashboard, not a matrix (Appendix F, Figure 91). P1209 stated, "All good information but they compete with each other." P1208 liked the idea, but not the display. She suggested, "Lift weights for

strain, smile/frown for quality, hand up for pause, throw out loudness. I could see all the

measures in one thing." This led to the multi-measure man (Appendix F, Figure 92).

Iteration 3

While there were more changes to individual measures, two big changes occurred in

iteration 3. First, the dynamic icons were simplified to static icons (Appendix G, Figure 93).

Second, the researcher changed how measures were presented. P1310 introduced the idea of a

similar layered structure for each measure. This incorporated multiple displays for each measure,

ordering them from the most "simple" to the more complex.

Icon Suggestions

Overall, there was little support for the dynamic icons. P1310 stated, "I feel the icon

should be consistent." P1311 liked the idea of dynamic icons, but P1312 was not in favor of the

dynamic icons. Therefore, icons were changed back to static images, as two of three participants

were not interested in them.

P1311 had an additional idea for the pause icon. To make it more iconic, it should have

"No words on icon." Suggested changing it to "Hands across mouth or something." Therefore,

the researcher tried a similar idea- drawing a zipper across the mouth to indicate no talking. This

was added to the other pause display so participants could choose which was better.

Loudness Suggestions

Participants were supportive of the two designs (Appendix G, Figures 94- 95). One

additional design was added by the researcher to assess whether participants might want to see

loudness go to zero during pauses because this hadn't been tested in prior iterations (Appendix G, Figure 96).

Pause Suggestions

Participants liked the idea of the tick mark display as a pop-out (Appendix F, Figure 83), so it was retained (Appendix G, Figure 97).

P1310 liked the idea of simplification. He didn't like the clocks or the lines. He suggested, "Maybe just put numbers for pauses." P1311 stated, "Bar graphs could be easier to look at." Therefore, the next iteration was a bar graph to indicate amount of time paused and numbers to represent the pause count (Appendix G, Figure 98).

Quality Suggestions

Overall, participants liked these displays, so they were all retained (Appendix G, Figures 99-101).

Strain Suggestions

The line graph was retained because this display was well liked by P1312, who preferred it and was especially complimentary on the double labeling of the axis (Appendix G, Figure 102).

The double line graph was not well liked. P1311 suggested changing it to a bar graph (Appendix G; Figure 104) and she described it as, "Have bar overlay- yellow for day and blue for night."

The third display was based on P1310's idea of making the similar layered structure across measures (Appendix G, Figure 103).

Multi-Measure Suggestions

The matrix was kept, but simplified based on feedback from P1310 (Appendix G, Figure 105). Two of three participants liked the idea of it, but P1311 preferred something simpler, like icons.

Iteration 4

In this iteration, the researcher introduced the idea of the layered display structure to participants, and they helped to further shape it.

Icon Suggestions

The biggest icon change was the decision between the two pause icons. P1415 commented that the "zipper on the teeth I don't like." P1414 expressed a preference for the stop sign. Therefore, the zippered mouth was not included in the next iteration, only the stop sign for pauses (Appendix H, Figure 106).

Loudness Suggestions

Participants were supportive of two designs from the prior iteration (Appendix H, Figures 108-109), but did not like the zero baseline display.

To match with the layered structure, participants suggested adding an "at a glance" version (Appendix H, Figure 107). P1413 suggested that it read, "Loudness- exceeded __dB for

__minutes." However, this was changed by the researcher to include the phrase "danger zone," which was a phrase used by multiple prior participants to describe what they felt the red color indicated.

Pause Suggestions

Participants liked the idea of the tick mark display as a pop-out (Appendix G, Figure 97), so it was retained (Appendix H, Figure 111).

P1413 remarked that the, "Average number helps me more than when it happened." This suggestion was used to create the simplified "at a glance" display for pauses (Appendix H, Figure 110).

P1414 commented that the, "Number on bars seems a little confusing" when looking at the pause displays. Therefore, a small multiple of bar graphs for each day was created without numbers to fit the layered structure (Appendix H, Figure 112). A zoomed in version of the bar graph was also created to match with the other zoomed in versions, such as loudness (Appendix H, Figure 113).

Quality Suggestions

Overall, participants liked these displays, so they were all retained (Appendix H, Figures 114-116) with one small change. P1413 was bothered by the fact that increasing strain was bad and increasing quality was good. Therefore, this participant suggested flipping the axes for one of the measures so that they would look visually similar.

Strain Suggestions

P1413 suggested putting both of the axis markers on the same axis. This was incorporated into Appendix H (Figures 118-119). The bottom figure is a simplified version Appendix G (Figure 102), without the numbers on the end, based on a suggestion from P1413 that the display was "too busy." This same simplification was used to remove the lines from the small multiple display by removing the lines. However, in this case, numbers were added based on a comment from P1415 that numbers would be helpful.

P1415 wanted a simplified view for the "at a glance," suggesting "Baseline in the middle, but for each event, indicate whether above or below the baseline." Therefore, the small multiple was simplified to arrows indicating whether the individual was above or below the baseline (Appendix H, Figure 117).

Multi-Measure Suggestions

The matrix was kept, but edited to reflect the changes in the other measures' displays (Appendix H, Figure 120).

Iteration 5 (Final) Results

Based on input from one repeat participant (P1208), the researcher finalized the displays, with some changes. These final images can be seen in Appendix I. Loudness and quality displays stayed the same (Figures 122 & 124), other than flipping the axes back for quality (this was confusing to P1208). P1208 was disappointed that the idea of multi-measure guy did not impress other participants. The researcher showed her the design, and this was used as a basis for completing the strain guy in the final displays. The zoomed in display for pauses was rejected

and the other displays were kept (Figure 123). P1208 looked at both current and prior displays

for strain before finalizing the design (Figure 125). The smiley faces were replaced by variations

of the strain icon to better differentiate between quality (smiley faces) and strain (weight lifter).

Summary of Results

Four measures were chosen by the participants (loudness, pauses, quality, strain). In

addition, participants recommended a layered structure of displays, rather than a single display,

for each of four measures. While the researcher kept the multi-measure display during all

iterations, about half of the participants expressed doubt on the usefulness or matrix design.

Therefore, because of the high level of disinterest in the multi-measure display, it was not used

as part of the feedback for Part 2.

Part 2 Methods

Using the feedback displays developed in Part 1, the researcher tested the feedback

display prototypes in Part 2. Potential target users completed vocal loading tasks (VLTs), tasks

designed to induce vocal fatigue, to determine whether objective feedback influenced later voice

production. Part 2 consisted of two phases: 1) laboratory testing (reading aloud for 15 minutes at

an elevated vocal intensity) and 2) field testing (classroom lectures).

Participants

Fourteen participants enrolled in Part 2.1, but only 13 participants met the inclusion

criteria of the study (6M, 7F; M=22.6 years, SD=9.15 years, Range=18-48 years). See Table 5

for demographic information on the participants. Only one participant, P2114, reported prior

speech therapy that was limited to a three-month time period in elementary school for an articulation disorder. P2101 failed the hearing screen at 4000 Hz in the right ear (able to hear at 25 dB), but this threshold was within an acceptable range for his age (Brant & Fozard, 1990), and so he was included in the study. On the other hand, P2108 failed the hearing screen at one frequency in one ear, and because individuals of her age should meet the 20 dB criteria, the researcher chose to exclude her from the study.

Participant ID	Age	Gender	Occupation/Major
P2101	48	M	Actor, radio volunteer
P2102	18	M	Undergraduate student, medicine
P2103	34	F	Graduate student, former elementary educator
P2104	18	F	Undergraduate student, neuroscience
P2105	18	F	Undergraduate student, engineering
P2106	19	M	Undergraduate student, business
P2107	18	M	Undergraduate student, criminal justice
P2108	18	F	---
P2109	18	M	Undergraduate student, engineering
P2110	19	M	Undergraduate student, packaging
P2111	25	F	Graduate student, media & information
P2112	20	F	Undergraduate student, wants to be professor
P2113	21	F	Undergraduate student, going to med school
P2114	18	F	Undergraduate student, biology; musical theater minor

Table 5: Part 2, Phase 1 Demographics.

An additional five participants enrolled in Part 2, Phase 2 (Part 2.2), and all participants met all the inclusion criteria (2M, 3F; M=42.6 years, SD=12.40 years, Range=33-62 years). See Table 6 for demographic information on the participants.

Participant ID	Age	Gender	Occupation
P2201	34	F	Course Instructor
P2202	36	M	Course Instructor
P2203	62	F	Course Instructor
P2204	48	F	Course Instructor
P2205	33	M	Course Instructor

Table 6: Part 2, Phase 2 Demographics.

Procedures

An overview of Part 2 can be seen in Figure 4. Prior to participation, all participants first completed an informed consent form approved by Michigan State University's Human Research Protection Program (HRPP). Participants completed three interview sessions (initial, midpoint, and final) and eight VLT sessions. The first three recording sessions (baseline recordings) did not include feedback, allowing the researcher to determine each participant's average baseline. After the midpoint interview, each participant completed five additional recording sessions (feedback recordings), receiving feedback at the beginning of each one.
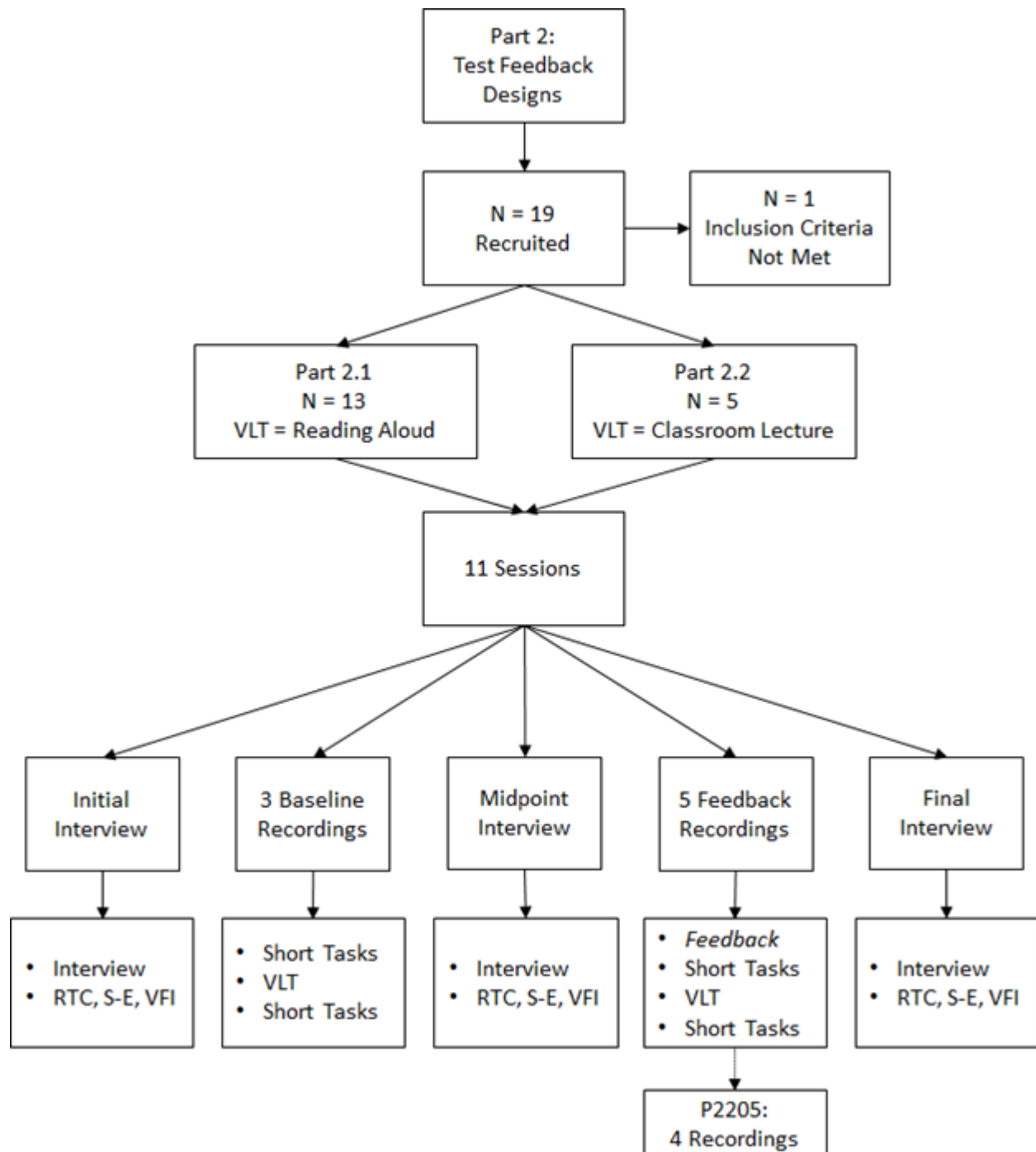
Figure 4: Outline of Part 2.

Interview Sessions

During each interview session, the participant completed the following:

1. Stage of change questionnaire: The URICA-VOICE (Teixeira et al., 2013) has been

   developed for use with the voice disorders population. Each participant's readiness to

   change was calculated from this assessment. Readiness to change is assessed by adding

   the average scores from the C, A, and M stages and subtracting the score from the PC

   stage (DiClemente et al., 2004; Teixeira et al., 2013).

2. Self-efficacy for voice change questionnaire: Self-efficacy was assessed using a modified

   version of a general self-efficacy scale (Lee, Hwang, Hawkins, & Pingree, 2008). The

   only modification was to the instructions: the word "voice" was substituted for the word

   "health."

3. Vocal Fatigue Index: Vocal fatigue was assessed along three factors using the VFI

   (Nanjundeswaran et al., 2015).

4. Semi-structured interviews: As in the semi-structured interviews for Part 1, the researcher

   has a series of printed questions, but asked follow-up questions when relevant to gain a

   deeper understanding of a participant's answers. The entire interview was recorded using

   two digital recorders to ensure no data loss (Roland R-05, Lake Stevens, WA, USA;

   TASCAM DR-40, Montebello, CA, USA). Some questions differed between the three

   interviews to better understand the participant's current and anticipated future voice use

   demands (initial; Appendix B), responses to voice monitoring without feedback

   (midpoint; Appendix J), and responses to feedback (final; Appendix J).

   a. Unique to the initial interview: Participants completed the general intake form

      (Appendix A), providing basic information about demographics and life factors

that can influence voice such as smoking history and caffeine intake. In addition,

the researcher gave a brief introduction to the study. The researcher described the

tasks that participants would complete and demonstrated how the equipment

worked.

b. <u>Unique to the midpoint interview</u>: The researcher provided a brief verbal

explanation of each measure and a basic introduction to each feedback display.

For example, for the voice quality displays, the researcher first provided

background on pitch strength as a measure of pitch saliency, and then briefly

stated that prior research indicated that this measure is a good predictor of listener

perception of voice quality. The researcher then went on to introduce the

participant to each type of quality display, and encouraged the participant to try to

walk her through the displays to assess understanding. Finally, the researcher

answered any questions the participant had about the feedback displays.

Recording Sessions

During the first recording session, the participant was fitted with an accelerometer

(VoxLog collar, Sonvox, Umeå, Sweden), which was attached to a Roland handheld recorder (R-

05, Lake Stevens, WA, USA). The accelerometer was attached to the front of the neck using

double-sided tape. The exact placement of the accelerometer varied participant to participant,

with placement dependent on where a strong signal could be detected by the accelerometer. In

general, the accelerometer was placed below the thyroid notch (Adam's apple). The VoxLog

collar consists of both an accelerometer and a microphone, with the accelerometer only recording

audio data up to 3 kHz. Recordings were dual channel (one channel each for accelerometer data

and microphone data). The participant wore the same dosimeter (accelerometer plus recorder) in all subsequent recording sessions.

For the feedback recordings, the researcher presented the displays with data from up to five prior recording sessions at a time on printed pages. Participants were allowed to ask clarification questions about the measures and displays (e.g., "What is quality again?", "Are increasing numbers better or worse?"), but they were not allowed to ask for the researcher's interpretation of the data trends. The researcher encouraged participants to "think aloud" while looking at the feedback to better understand how participants (potential users) interpreted the feedback with limited outside guidance.

Recording Sessions- Part 2.1

VLTs were all completed in the same single-walled, sound-treated booth while wearing the dosimeter. The configuration of the booth can be seen in Figure 5. The VLT consisted of 15 minutes of reading aloud from a novel presented in electronic form on a tablet (Charlotte's Web by E. B. White) while seated. This duration was chosen by the researcher because is less than the 35 minute safety limit suggested for a continuous reading task (Titze et al., 2003). In addition, the researcher instructed participants to "Read aloud as though you are reading to a classroom and want to be heard by all the children," which are the same instructions given in a prior study to elicit louder speech (Hunter et al., 2015). Loud speech is one type of task that is common to induce vocal fatigue in participants (Solomon, 2008). This instruction allowed the participants in Part 2.1 to approximate the vocal effort of the course instructors, although for a shorter time. Participants were instructed to read at their own pace. The researcher informed participants they had the option to end any VLT early if their voice became very tired or if they needed to leave

the booth before the end of the timed task (such as to use the restroom). As a precaution, if a participant's initial vocal fatigue was 7 or higher (none were), the researcher would cancel the session for that day and reschedule for another day. Participants were told to cancel if sick.

The researcher instructed participants to try to maintain a similar vocal loudness throughout the task. The target dB level was established during the short tasks prior to the VLT. Participants counted from 1-5 at a comfortable and then a loud level. If the level became too low (no peaks occurring above the average comfortable counting level), a lamp would come on in the room, indicating that the participant should increase their volume.
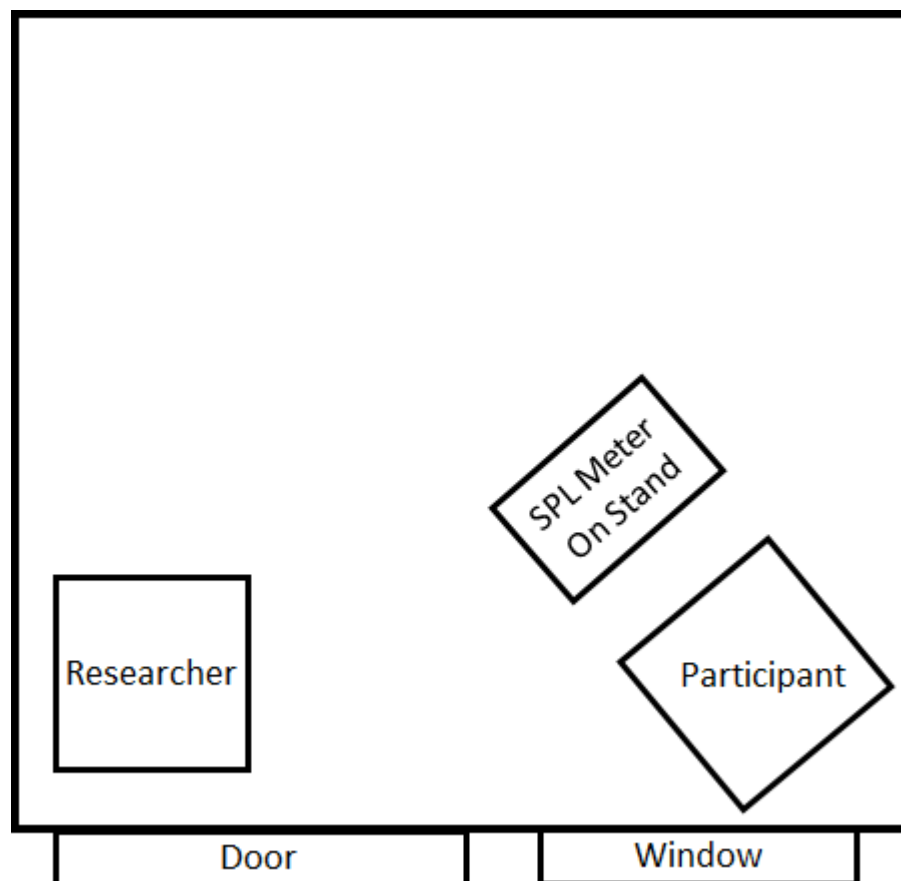
Figure 5: Sound-treated booth configuration for Part 2.1. The two squares indicate the location of chairs for the participant during the session and the researcher during the presentation of the feedback.

Recording Sessions- Part 2.2

Part 2.2 had the same structure as Phase 1, but the vocal loading task consisted of recordings during classroom lectures. During the VLT, no additional instructions to maintain a particular vocal loudness were given. The recording policy was as follows: if the instructor held class and was comfortable with the recording that day, the recording was made. If an instructor was sick and did not want to record that day, their wishes were respected. If an instructor informed the researcher that there was an exam in class on a particular day and there would be limited instructor talking, the recording did not occur that day. If the instructor did not hold class on a particular day, the recording did not occur that day.

Due to the real-world nature of the Part 2.2 recordings, additional precautions were taken. During the first interview session, the researcher made participants aware that not only their voices would be recorded, but the voices of their students might also be recorded (on the microphone recordings). The researcher instructed participants to explain this to students, giving students the option to let the instructor know if recordings made them uncomfortable. Fortunately, no student concerns were reported to the researcher. As an extra precaution, participants were taught how to pause the recorder, and the researcher informed participants that they could turn off the recorder at any time if they felt there were any privacy issues, and to let the researcher know if anything sensitive was recorded so that it could be deleted.

All instructors taught on Mondays and Wednesdays (which was not planned by the researcher). Due to the real-world nature of the study, the course length and amount of lecture material varied (Table 7). Exact course length varied week to week even for the same class meeting, so the lengths reported are the assigned lengths in the course directory.

Participant	Monday Class Length (hrs)	Wednesday Class Length (hrs)	Notes
P2201	1.5	3.0	2 different classes
P2202	3.0	3.0	2 different classes: M is more student presentations, *W is more lecture based
P2203	1.5	1.5	Same class
P2204	1.0	1.0	Same class
P2205	3.0	3.0	2 different classes: toward end of semester, more voice use in W class

Table 7: Basic course structure, by instructor.

Feedback

The feedback presented to participants resulted from Part 1 of the study. Based on the findings from Part 1, the measures for feedback were (with names used for feedback in parentheses): intensity (loudness), pitch strength (quality), first moment specific loudness (strain), and pauses (pause frequency and pause duration). However, because the researcher instructed participants to artificially increase their vocal intensity in Part 2.1 (instructions to "read as though reading to a classroom"), feedback on loudness was not provided. Therefore, participants in Part 2.1 received feedback on three measures (quality, strain, pauses), and participants in Part 2.2 received feedback on four measures (quality, strain, pauses, loudness).

For the recordings with feedback, participants received feedback information prior to VLTs. An example of the feedback, with explanations of each display, can be seen in Appendix K (Part 2.1) and Appendix L (Part 2.2). Feedback was presented on paper, one display at a time, but participants could request to go back to look at prior displays or look at more than one display at a time if they felt that a comparison was helpful. The researcher randomized the order of feedback presentation across participants to ensure that there were no order-related effects.

Part 2 Analysis

Aim 1

In Part 2, the researcher sought to determine what further design requirements could be incorporated into future iterations of the feedback displays. Therefore, midpoint and final interview responses related to occupational voice user needs, feedback design, and feedback usefulness were transcribed, coded, and analyzed. Coding was based in grounded theory (Corbin & Strauss, 2008). After each interview was transcribed, the researcher used open coding to extract the meaningful data from the interview. Open coding relies on the data to generate the codes, rather than assigning pre-conceived codes to the data. For example, if a participant remarked, "I like the strain and quality measures", two codes may be assigned: "likes strain measure" and "likes quality measure."

After all coding was completed, the researcher used affinity diagramming to identify emerging themes from the data (Beyer & Holtzblatt, 1998). Affinity diagramming implements a bottom-up approach to data analysis, allowing themes (code groupings) to emerge based on the data presented. The most common themes, representing codes from multiple potential users, can then be used to inform future design iterations.

Aim 2

Aim 2: To identify changes in voice behavior management after receiving feedback. The researcher used open coding for midpoint and final interview results to provide further insight on intuitive behavior changes and changes facilitated by the feedback. In addition to identifying themes from the interviews, it was hypothesized that active engagement in changing voice

production will manifest as improvements in: RTC (increase), S-E (increase), and vocal fatigue (decrease for dimensions 1 and 2, increase for dimension 3).

To address this hypothesis, the researcher completed a mixed analysis of variance (ANOVA). The researcher chose this analysis because it compares multiple measures while controlling for the other variables. This analysis was using repeated measures, but due to the addition of a between-subjects factor (gender), the correct term for the analysis is a mixed ANOVA. In this analysis, there were two within-subjects factors: time (initial, midpoint, and final interviews) and measure (RTC, S-E, and VFI scores). The researcher entered each of the three factors of the VFI separately, leading to a total of five measures in the analysis. This is in agreement with (Nanjundeswaran et al., 2015), where it is recommended that each VFI factor be reported separately. In addition, gender was the between-subjects factor to determine whether gender influenced participant responses. Statistical analysis was completed using IBM SPSS software (IBM SPSS Statistics for Windows, Version 23.0. IBM Corp.; Armonk, NY, USA).

Part 2 Results and Discussion

Aim 1 Interview Results and Discussion

A total of 1064 unique codes were generated from the interview data, with some repetitions of codes when appropriate. Of those codes, 430 were classified into themes related to Aim 1. The three most common themes are reported, comprising 323 unique codes (Table 8). They are: positive comments on current feedback displays (66 codes), occupational voice user needs (152 codes), and recommended feedback display improvements (101 codes). These three themes encompass the most important commentary and recommendations of the participants for

the feedback displays. Detailed descriptions and examples of quotes fitting these themes are

included below.

Theme	Unique Codes
Positive Comments on Current Feedback Displays (66 codes)	
1. Displays are user-friendly	25
2. Measures are helpful and should stay included in feedback	41
Occupational Voice User needs (152 codes)	
1. Clearer definitions of measures are needed	66
2. Strategies for improving the voice based on feedback are needed	64
3. The system should be adaptable for a range of user needs	24
Recommended Feedback Display Improvements (101 codes)	
1. Users should be able to include notes and labels in data	16
2. Displays should show relative trends across days	24
3. Additional feature suggestions	61

Table 8: Emergent themes related to Aim 1.

Theme 1: Positive Comments on Current Feedback Displays

It is promising to note that some features of the current displays are felt to be user-

friendly and helpful. While participants provided many helpful suggestions for changes in future

iterations of the feedback displays, some elements have been identified as positive features that

should be preserved in future iterations.

Sub-Theme 1: Displays are user-friendly

Multiple participants expressed that the feedback was simple and easy to read in both Part

2.1 and Part 2.2. For example, P2106 stated, "I thought it was really easy to read, very simple,"

and he added, "The symbols do a pretty good job representing each quality [measure]." P2101

reported there was "nothing wrong with how it [the feedback] was presented." P2205 highlighted

strain as a measure with good displays: "It shows the changes intuitively." Overall, P2205

remarked, "Already, it's kind of a practical system." In summary, while participants did identify weaknesses in the current feedback displays, participants felt that one strength of the current displays is the straightforward way that they are presented, and this presentation manner should be maintained in future iterations.

Sub-Theme 2: Measures are helpful and should stay included in feedback

All four measures were identified as being helpful by at least one individual. Of the four measures, strain was mentioned the most often (10 instances), then quality (4 instances), pauses (3 instances), and loudness (2 instances). Some participants felt that there was not just one most important measure. P2105 commented that strain and quality were the most helpful, while P2111 felt that strain and pauses were the most helpful. P2104 remarked, "All [measures were] about the same in helpfulness." Some student participants in Part 2.1 provided insight into how they felt occupational voice users would interpret the measures. For example, P2114 commented, "I definitely think that strain is something that's very important to look at, especially if you're a professor talking all day". This corresponded with a comment from a participant in Part 2.2, P2204: "It's nice to know how much strain is involved in the voice." While strain was identified as the most helpful measure, participant commentary suggests that all measures provide some level of usefulness and may be appropriate to maintain in future iterations.

Theme 2: Occupational Voice User Needs

The core feature of user-centered design is involving the end user in all stages of the design process to the greatest extent possible. While occupational voice users and students who are future occupational voice users were included in Part 1 of the study, they did not actually use

the feedback they helped to create. The additional occupational voice user needs uncovered by

the participants in Part 2 included: clearer definitions of measures need to be provided, objective

feedback needs to be accompanied by strategies to improve voice use, and individual user needs

demand greater flexibility in the system.

Sub-Theme 1: Clearer definitions of measures are needed

While some participants felt that the displays were easy to use, many participants

struggled with understanding the measures featured in the displays. A simple explanation of each

measure was provided in the midpoint interview when the feedback displays were introduced,

and participants were able to ask for clarification on measures in feedback recording sessions,

but this was not perceived to be enough support for understanding the objective measures. The

measures most frequently identified as needing more detailed explanations were quality (16

instances) and strain (15 instances). While to a lesser degree, pauses (5 instances) and loudness

(2 instances) were also identified by participants as needing more detailed explanations. As

stated by P2107, "If I was going to get an app like that on my phone, I wouldn't fully understand

[the measures] unless there was a description of it."

This lack of an understanding of the measures led to confusion, as explained by P2112,"

But I had no idea why my quality varied and why my pauses varied." P2111 reflected, "I

understand that I want to improve my quality even though I have no idea what is the higher

quality." Some participants, such as P2201, even remarked that not having clear definitions led

to confusion between measures when she stated, "More clear distinction would be helpful."

P2103 felt that not just the measures themselves needed better explanations, but also, "Specific

strategies or helping to make links across the three pieces of feedback" would benefit users. This

confusion and need for more explicit definitions of measures and relations between measures was summed up by P2203, "If I was a specialist, I could look at what it is [feedback measures] and I could do something about it, but as a novice, I'm not invested in this in the same way." With many of the participants expressing the need for a better understanding of the measures, this suggests that a simple explanation of each measure is not enough and future design iterations need to come up with a more explicit option for definitions, such as optional viewing when needed as suggested by P2107, "something you can click on, or if you click on the thumb and you can see your quality and a description up at the top and individual days underneath."

Sub-Theme 2: Strategies for improving the voice based on feedback are needed

While physical activity trackers, in their simplest form, provide a daily step count for individuals to track their activity over time, basic reporting of measures is not sufficient for PVM. P2102 nicely described his interpretation of the difference between a physical activity tracker and voice monitor, "[For a voice monitor] Give interpretation and recommendation, the most important part. If people buy this, it's what they'll be looking for. It moves past what a physical activity tracker is." This desire for providing strategies for improving the voice as part of feedback was echoed by most participants. P2103 stated, "To me, it was informative feedback, but not very educative. The feedback didn't help me to know what to do." She later added, "In education, we not only give feedback on performance but on how to improve." P2202 felt the same, "The feedback is interesting, but it is coaching of things you can do and try that I think would be helpful also." P2204 also highlighted the importance of including strategies by describing an alternative scenario, "Otherwise, every individual will make his or her own decisions and you will have absolutely zero control over what they do." Based on this

information from participants, it is clear that feedback from voice monitoring needs two components: 1) the measures themselves and 2) suggestions for how to improve voice use based on the measures. The exact nature of these suggestions, such as the depth of explanation needed, should be explored in future studies.

Sub-Theme 3: The system should be adaptable for a range of user needs

When looking at occupational voice users from a vocal health perspective, the focus is on reducing risk factors for developing voice disorders. As mentioned previously, these include decreasing loudness and decreasing the time spent talking (increasing pauses). However, just as voice therapy is tailored to individual client needs, participants suggested that PVM also needs the flexibility to adapt to specific user needs. Again, this differs from a simple version of an activity tracker, where all users are assigned the same basic goal.

P2203 discussed this idea in her final interview: "I think it also speaks to the function of the person talking. If I'm talking on the radio, it's a different kind of thing. If I'm talking to a classroom where I have an audience and I have response and I'm trying to convey certain things, then my big picture of how my voice is being used is a different thing. If you're talking because you're doing audio books, that's a different thing too. Different in how you use your voice, and different in what's important."

P2103 discussed the PVM system in relation to the specific demands placed on teachers: "Losing voice is a constant problem for teachers. Losing your voice is especially a problem for new teachers. There's a real need for it [PVM]." Despite the need, P2103 went on to discuss the difficulty of getting teachers to adopt such a device with the demands they face on a regular basis

as part of their occupation: "This would be one more new thing that teachers have to think about buying into."

Not only were varying occupational voice user needs identified in the abstract, but study participants suggested varying goals within the study. P2101, P2104, P2111, and P2113 all identified themselves as being "quiet talkers." However, these individuals reported the VLT (reading at an elevated vocal intensity) helped them to speak louder in their daily lives. P2101 stated, "I speak louder now." P2111 integrated increased vocal intensity and pause time in her teaching, "So I trade off- I pause more but become louder for more people to hear me." On the other hand, P2201, P2110, and P2114 reported being "loud talkers" trying to speak quieter as a result of the study. P2101 reflected, "I try to be more careful with my voice." P2110 said, "If I don't have to talk that loud, then I usually don't" in the midpoint interview (before feedback), but reported no change in vocal habits outside the study in the final interview. This discrepancy suggests that either some changes as a result of voice monitoring may be more short-term, or maybe that this change was integrated into his vocal behavior and was no longer seen as a change from the study.

These comments from participants provide an important consideration for future development of PVM: it is not enough to simply suggest voice conservation to reduce the risk of disorders. Individual goals need to play some role in the feedback. Further investigation is needed to better understand the breadth of these different needs, and to determine whether any common ground can be identified.

Theme 3: Recommended Feedback Display Improvements

Participants felt that there were a number of ways to improve the feedback displays to increase the functionality of PVM. These included adding notes and data labels, focusing on relative data trends rather than absolute values, altering symbols and baseline values, and adding real-time alerts to the user.

Sub-Theme 1: Users should be able to include notes and labels in data

One issue commented on by multiple participants during the study and addressed again in the final interviews was the difficulty with delayed recall of vocal behavior days after it occurred. The inability to analyze the voice recordings online was a limitation of the current study, but based on participant comments, difficulty with delayed recall would be present even with real-time analyses when users try to compare voice behavior across days. P2205 stated, "The voice tone is related to the context- I want more cues for context."

P2202 suggested, "It would be really cool to go back through and label time periods… Part of the reason for the study is not just providing information but helping to make changes. I think the chunking would actually help with that." Or, he suggested a more automated way of doing this: "I'd love to line it up with my syllabus because that would really generally remind me of what I was talking about that day." For non-teachers, P2112 proposed, "They could have a little calendar and they can go back and look at 'oh, this accounted for it'." In addition to including context about what was going on in the feedback, participants also wanted a notes function. For example, P2102 felt that users would like a way for "Adding your own interpretations." P2202 also wanted a "'Notes to self' field for each day." While participants

wanted both feedback and tips for how to improve voice production from a monitoring system, they also want to add context to better understand their voices.

Sub-Theme 2: Displays should show relative trends across days

The desire for relative trends is also partially related to difficulty with delayed recall of voice use. As stated by P2202, "Because of the time delay, it was easier to think in large time chunks." It is also related to a personal preference, as expressed by P2204, "My preference is to see the small clearly and the big picture." P2205 echoed this sentiment, "Relative trends or changes were helpful for me." P2111 took this idea a step further: "For the different weeks, probably my voice will change a little bit. Sometimes I'm really tired, sometimes I'm really energetic. So doesn't mean anything- or probably an average value of one month or one semester, that probably make more sense."

For the current study, the feedback was scaled to show a short amount of time (15 minutes to 3 hours) across days in a detailed form. However, the amount of detail in displays comparing the voice over full days needs to be investigated further. Based on the opinions of some study participants, a high level of detail may not be as meaningful to users as general trends across days.

Sub-Theme 3: Additional feature suggestions

Participants made many suggestions for future PVM development in addition to those described above. Some additional suggestions are included to demonstrate the breadth of these recommendations, and the need for further research before PVM is ready for widespread adoption.

Some participants suggested alterations to the symbols used in feedback. Specifically, suggestions were made for the changing symbols along the y-axis for both quality and strain. P2201 was concerned that the baseline face is "…sometimes perceived as a frowny face for some reason," and suggested including a frowning face as a contrast to the baseline face to help remind users. P2205 was concerned about the strain guy stating, "Sometimes it looks not that bad because it's working harder." This participant was referring to the fact that the symbol for more strain looks like it's working harder at lifting weights, which can be good for one's health. P2201 felt that simplification of the strain image could help, stating: "the face with a little bit of effort, and the face with the lines" rather than picturing the weights. P2109 had the opposite reaction: "The stick person might be simpler- Maybe just a barbell."

P2202 was concerned about the changing baseline for loudness (baseline was based on the average dB level by day). This was different than strain and quality, which had the same baseline each time. He felt it "would be nice if they all had same baseline so you could compare across days." P2112 suggested adding a baseline value for pauses, "It'd be nice to have a baseline for the pauses, since there's the baseline for the other two."

Participants also want alerts built in with varying functionality. P2205 suggested a "nudge" feature: "I was thinking 5 minutes before the class, if the system can have access to my calendar, 'your class is coming. In your last class you did this, and this, and this,' in my Apple watch. And then 'why don't you…'" Based on voice use, P2106 wants the PVM to "Give you alerts to say you're at high risk for voice problems right now." P2111 wants more specific real-time suggestions, such as: "I would suggest you have a rest, or you should have a pause" when the PVM detects that the speaker has been talking too long. P2205 offered that loudness "could be volume as a trigger or a cue for the adjustment."

Aim 1 Interview Results Summary

The interview results were very informative for considerations in the future development of PVM. The main findings included some positive reviews of current features, including the simple display organization and helpfulness of the current measures. Participants also suggested many ways to improve PVM to appeal to a wider audience by allowing flexibility for varying user goals and providing greater support through clearer measure definitions and suggestions for how to improve one's voice production based on the current feedback. Finally, participants suggested adding a level of customization in the displays so that users could add notes and labels to better understand the context of the data, focusing on relative trends in the data, changes to improve interpretation of the displays, and alerts when behavior suggests an increased risk for voice disorders.

Aim 2 Interview Results and Discussion

A total of 1064 unique codes were generated from the interview data, with some repetitions of codes when appropriate. Of those codes, 421 were classified into themes related to Aim 2. The three most common themes are reported, comprising 343 unique codes (Table 9). These three themes encompass the overall picture of change for these participants: changes in action, thinking, and lack of change. Detailed descriptions and examples of quotes fitting these themes are included below.

Theme 1: Reported Behavioral Changes

Participants reported a variety of behavioral changes in the interviews. These included changes to try to help the voice, changes stemming from knowledge of being monitored, and

some specific changes reported due to the nature of the VLT in Part 2.1. These are described in detail below.

Theme	Unique Codes
Reported Behavioral Changes (156 codes)	
3. Active changes in vocal behavior due to increased awareness and feedback	79
4. Changes in vocal behavior due to monitoring	17
5. Task specific voice changes in Part 2.1	60
Increased Awareness of Voice (138 codes)	
4. Interpretations of feedback	106
5. Learning about own vocal fatigue and risk of voice problems	32
No Observed Changes (49 codes)	
4. No conscious behavioral change	35
5. No change observed in feedback measures	14

Table 9: Emergent themes related to Aim 2.

Sub-Theme 1: Active changes in vocal behavior due to increased awareness and feedback

Participants reported a range of behavioral changes resulting from participation in the study. The most common change was increasing pauses (24 unique codes). P2111 reported increasing her pausing behavior in her teaching outside of the study, "When I find it's (lecturing) boring, I probably speak more fast. So I try to commit to thinking 'I need to slow down, I need to have pause.'" P2202 reported, "The place that I noticed the most change was how much [I talked]." He continued by revealing a strategy he used to increase pausing (voice rest), "When I was starting to notice things, I was trying to get the students to talk more."

Changing vocal loudness with the second most common change (22 unique codes), but the direction of change varied between individuals. This was an unexpected finding given that the purpose of PVM was to reduce the risk of voice disorders, and one way of doing that is reducing loudness. P2114, who reported being a "loud talker," followed the expected direction:

"There are periods where I'll talk quieter if I'm like 'Oh, my voice isn't feeling the best today.'" P2110, another "loud talker," also reported, "I can be quieter [than before the study]." On the other hand, P2113, a "quiet talker," stated, "Outside of the study I might have started talking a little bit louder."

Other changes were reported, such as P2111, a "quiet talker," who stated that she now balances her vocal loudness and pauses in her teaching: "So I trade off- I pause more but become louder for more people to hear me. I think that will make more clear to the students." P2205 reported that during the baseline recordings he "tried a little bit to find a voice that made my throat comfortable," including changing the pitch of his voice.

Sub-Theme 2: Changes in vocal behavior due to monitoring

Not all voice changes were due to participants trying to change their voice to help voice production. Some of these changes were made due to the monitoring, many of which became reduced with continued monitoring. These changes were reported by participants in Part 2.1 and Part 2.2. However, these changes probably affected how participants thought about their voices in the study, and may overemphasize changes that were seen. For example, P2201 reported at the midpoint interview that she "tried to make sure you had enough information, so I tried to talk more." P2109 and P2112 reported focusing on enunciation, as evidenced by P2112's comment: "I think I focused more on how I was enunciating- I focused on saying every word." P2114 stated, "I notice that I eventually got louder over time [within the study]."

Sub-Theme 3: Task specific voice changes in Part 2.1

Some behavior changes that were noted by participants were specific to the reading task in Part 2.1. Again, these may inflate some of the participant-reported findings in the study, especially if they made these changes sound more generalized in the interview. Some of the more common reports included talking louder than usual (10 instances) and talking more than usual (7 instances). For example, P2104 stated, "And I know I have to talk louder [in the VLT]" and "I talk more than I usually do [in the VLT]." Other changes were reported, such as P2114 stating, "I did notice that while reading Charlotte's Web I had specific voices for each of the characters. Sometimes one of them [the voices] would bother me so I would change it."

Theme 2: Increased Awareness of Voice

Participants reflected on their change in mindset when it comes to voice. Participants shared their interpretations of the feedback, as well as increased awareness of vocal fatigue and risk of voice problems. These sub-themes are further explained below.

Sub-Theme 1: Interpretations of feedback

Because there was little direction given to participants in how to interpret the feedback, insights into the thought process of these individuals is critical. This information can be used in future iterations to shape the presentation of data, and may also be of use to other voice care professionals when designing treatment programs.

First, 16 unique codes described disagreements between what participants saw in the feedback and their expectations. For example, P2106 felt he had a good grasp on the relationship between measures until "I had the day that completely threw my theory off, but that might have

been a fluke day." P2110 discussed the unpredictable nature of the quality measure: "Some of its random, it feels like sometimes. Like when the quality is jumping up and down." P2113 felt there was a disconnect between her perception of vocal fatigue and strain: "I feel for the fatigue one [lowest on the scale], I have no idea what I'm going to get [for strain]."

While there was some disagreement with the feedback, some comments suggested a level of agreement or a connection between the participant and the feedback. P2202 commented, "This was more useful [pause bar graphs] because it's over time especially because you can see the structure of my class." P2205 saw the feedback as a way to interact with his voice on a different level: "An avatar is reacting to my performance, so making this guy calm was the objective. It is showing myself, it is showing my throat. I need to take care of myself."

Another interesting finding was the level of trust in the feedback. While the feedback did not conform to some participants' ideas of how it should work, some participants indicated a high level of trust in the feedback. P2109, who noticed that his strain often decreased from the beginning of a session to the end while his fatigue increased commented, "I paid less attention to how my voice felt. I'm not talking for that long, and it will be fine, according to the data that I've seen." This was counter to what was anticipated in the study, with the goal being to increase voice awareness rather than relying on feedback to tell you what is happening. However, the relationship described by this individual, of decreasing strain over time, was seen by multiple participants and may be due to a warm-up effect (Vilkman, Lauri, Alku, Sala, & Sihvo, 1999) rather than simply showing that strain decreases over time with use. Possibly repeating this study with a longer VLT may lead to different results.

Sub-Theme 2: Learning about own vocal fatigue and risk of voice problems

One common area that participants expressed increased awareness was in vocal fatigue. P2202 was especially excited about noticing gradations in fatigue, which he had not noticed previously: "Participating in the study versus not, I was noticing those indications [of fatigue]. I was noticing a lot of differences rather than all of a sudden noticing that my voice was really, really tired. Now I can notice more gradations in between which has been really cool." P2106 also discussed his increased awareness of vocal fatigue: "I became more aware of when I needed to give my voice a rest or when it was getting more effortful to produce." P2201 took this a step further, expressing that before the study, "I never even thought there was such a thing as a voice problem," but that now she is much more careful of how she uses her voice.

Theme 3: No Observed Changes

Sub-Theme 1: No conscious behavioral change

While many participants reported changes in their vocal behavior, either task specific or generalized, other participants reported a lack of behavioral changes. For example, P2204 stated, "I need to communicate, and I started where I ended in the sense I talked in the same fashion." P2203 also did not feel that she changed her voice as part of the study, but rather, she felt that: "from my point of view, I'm just participating." Other participants changed some things, but not others. For example, P2201 remarked on her loudness: "I don't think much changed in terms of loudness of my voice."

Sub-Theme 2: No change observed in feedback measures

Some participants commented that they saw little to no change in the feedback measures over the course of the study. For example, P2107 stated, "I didn't think the vocal strain was as helpful because I didn't feel like it changed all that much. But then again, neither did the other two." P2112 attributed this lack of perceived change as possibly stemming from the scaling of the feedback, remarking, "It'd be nice if our baseline was a little more tight so that we could still see change." However, another reason why there may have been little perceived change stems from the design of the experiment. All participants reported no history of or current voice disorders requiring medical treatment. Therefore, all their voices were relatively normal, and this lack of change may stem in part from not having voice problems. It may be informative to try a similar task with individuals with voice disorders to see if greater changes are observed.

Aim 2 Interview Results Summary

Overall, participants reported a wide spectrum of changes in the study. These changes ranged from a lack of change to increased awareness (changes in thinking) to actual behavior changes. Within each of these categories, the reports varied. For example, some individuals reported an overall lack of change in behavior, whereas others reported a lack of change for specific measures and changes in others. Increased awareness ranged from increased awareness of different aspects of the voice through interpretation of the feedback to increased awareness of the presence of vocal fatigue and risk of voice disorders. Finally, behavior changes were reported. Some changes were generalized to a broader context while others were limited to the VLT.

Interview Results Limitations

There were some limitations to Part 2 that may have influenced the interview results. First, participants in Part 2.1 were predominantly students who had limited prior experience as occupational voice users. Therefore, some of their comments were based on what they think would relate to occupational voice users. In addition, the controlled reading task may not have offered enough similarity to a real-world environment to facilitate carry over of voice use changes, and their insights may have changed with a more real-world VLT.

Some of the changes reported by participants in Part 2.1 were related to the task itself rather than actual behavior change, such as P2113's reflection that the way she talked in the VLT was "Definitely louder than I normally talk." When participants clarified these statements as being specific to the demands of the task, they were coded this way, but not all participants provided this level of detail. Therefore, it is possible that the level of perceived change in Part 2.1 may be artificially inflated.

It is also possible that the length and structure of the VLT in Part 2.1 may be too short to elicit the expected changes. It is possible that the task was too short for some participants to experience fatigue (reported by P2101), and for others, it may lead to a warm-up effect rather than a fatiguing effect. P2110 commented that it would be interesting to compare vocal performance between this length of task and a longer VLT. P2107 was interested in seeing what effect different reading material (with more emotional content) may have on the voice.

Part 2.2 was the field test, and involved recording in participant's natural environment. Therefore, it is more likely that the changes reported are from participating in the study rather than an artifact of the task itself. However, the scope of the recordings is still limited. Other vocally demanding tasks, such as teaching other courses and attending meetings, were not

recorded. Therefore, participants were not given a more holistic picture of their day, but rather, a snapshot.

Aim 2 Statistical Analysis

Out of all the questionnaires, only one missing data point was identified (for the final interview S-E questionnaire for P2204). However, because the rest of the questionnaire was answered, this data was still used with the understanding that this individual's score was potentially lower (e.g., they could have scored the question as "0") due to the missing data. Because all questionnaire scores were used, a total of 13 scores were analyzed (one per participant) for each measure at each interview in Part 2.1, and 5 scores were analyzed (one per participant) for each measure at each interview in Part 2.2.

It was hypothesized that active engagement in behavior change would manifest as improvements in: RTC (increase), S-E (increase), and vocal fatigue. The improvements in vocal fatigue were defined as changes in one or more of the following of the VFI (consistent with Nanjundeswaran et al., 2015): a decrease in Factor 1, a decrease in Factor 2, and/or an increase in Factor 3. For changes in individual results in Part 2.1 across the three interviews, see Appendix M, Appendix N, and Appendix O for RTC, S-E, and VFI scores respectively. For changes in individual results in Part 2.2 across the three interviews, see Appendix P, Appendix Q, and Appendix R for RTC, S-E, and VFI scores respectively.

Part 2.1

The assumption for sphericity was met for all variables but "measure." Therefore, Greenhouse-Geisser corrections were used for all Part 2.1 questionnaire statistics for consistency. The results of the mixed ANOVA are reported below.

Main Effects

There was a significant main effect of time, $F(1.947, 21.417) = 6.299$, $p = .007$, $\eta^2 = .364$. The mean score across questionnaires was significantly lower in the initial interview (M= 7.798, SD= 4.495) than either the midpoint interview (M= 9.081, SD= 3.237) or the final interview (M= 8.902, SD= 4.335), $p = .026$ and .031, respectively. No significant difference was found between the midpoint and final interviews. This finding is in the hypothesized direction for three of the five measures (RTC, S-E, and VFI Factor 3). However, it is of interest to note that the trend is for these measures to increase prior to the feedback, suggesting an effect of monitoring. Trends are explored in more detail below. Corresponding individual results can be seen in Appendices M-O (Figures 140-142).

For RTC, this trend of increasing and then decreasing was seen in 9 of the 13 participants (69.2%). This finding provides support for Aim 2: behavior change measures will show improvement. The trend suggests that most participants increased their readiness to change, likely due to actively engaging in the behavior change process. Even more interesting, all 9 participants had a higher RTC at the midpoint interview (being monitored with no feedback) that declined (although still above initial) in the final interview. Future work should further explore the reason for this great increase in the monitoring stage, but decrease after feedback.

For S-E, the same trend was seen with 6 participants (46.1%). In addition. 8 participants (61.5%) showed at least an increase from the initial to the final interview. These findings suggest that over half of participants experienced an increase in self-efficacy over the course of the study, which is again in support of Aim 2. Again, future work should further explore the reason for the greater increase in S-E during monitoring alone (versus monitoring with feedback).

For VFI scores, the findings were mixed. On VFI Factor 1, 11 participants (84.6%) followed the trend seen for RTC and S-E. This is counter to the hypothesis in Aim 2 because increasing scores on Factor 1 indicates greater vocal fatigue. However, one possible explanation for this increase may be that because most of these participants were students with limited prior experience as an occupational voice user, engagement in the VLT may have increased their awareness of vocal fatigue, leading to increased scores. In this case, the decrease in the final interview suggests at least some reduction in perceived fatigue due to engagement in PVM.

VFI Factors 2 and 3 showed much weaker adherence to the overall trend, with 4 participants (30.8%) and 3 participants (23.1%), respectively showing the trend.

There was also a significant main effect of measure, $F(2.159, 23.754) = 32.402$, $p < .001$, $\eta^2 = .747$. Almost all measure pairings were statistically significantly different at or below $p= .043$ with a few exceptions. The exceptions were: RTC (M= 4.621, SD= 2.153) and VFI Factor 2 (M= 4.513, SD= 2.999) at $p= 1.000$; S-E (M= 9.744, SD= 3.470) and VFI Factor 1 (M= 14.487, SD= 4.541) at $p= 0.71$; and S-E and VFI Factor 3 (M= 9.641, SD= 2.311) at $p= 1.000$. This main effect was less meaningful because the scaling of the questionnaires and the underlying constructs they measured were different, and these differences probably contributed to the significant differences. However, the findings are consistent with the three factor model of vocal fatigue because all three factors were statistically significantly different.

Finally, gender had a non-significant difference, $F(1,11) = .045$, $p = .835$, $\eta^2 = .004$.

Interactions

There was a non-significant interaction of time and measure, $F(4.024, 44.267) = 2.444$, $p = .060$, $\eta^2 = .182$. Despite the non-significance of this measure with the Greenhouse-Geisser correction, this nearly significant result warranted further investigation to see the emerging trends. (If a Huynh-Feldt correction had been used instead, this result would have been significant at $p = .024$.)

Individual comparisons can be seen in Tables 10-12 for the initial, midpoint, and final interviews, respectively. In these tables, only the statistically significant differences are shown. Some interesting findings emerge when looking at these results. The reported means are difference scores- the difference of the row variable minus the column variable (e.g., RTC minus S-E). These differences show, for example, that S-E scores were higher than RTC scores (on average) in the initial interview. Again, while these differences were significant, these findings are of less interest to the overall purpose of the study.

Initial Interview	S-E	VFI Factor 1	VFI Factor 2	VFI Factor 3
RTC	-4.94 (3.78) .006	-8.13 (4.96) .001		-5.13 (3.23) .001
S-E				
VFI Factor 1			8.18 (4.59) <.001	
VFI Factor 2				-5.18 (2.97) .001

Table 10: Comparison of Part 2.1 scores for the initial interview. Reported are the mean difference in measures across participants (standard deviation) and p-value.

Midpoint Interview	S-E	VFI Factor 1	VFI Factor 2	VFI Factor 3
RTC	-4.69 (3.33) .004	-10.91 (3.06) <.001		-5.16 (3.21) .001
S-E		-6.23 (5.53) .019		
VFI Factor 1			10.69 (2.58) <.001	5.75 (3.72) .002
VFI Factor 2				-4.94 (3.66) .005

Table 11: Comparison of Part 2.1 scores for the midpoint interview. Reported are the mean difference in measures across participants (standard deviation) and p-value.

Final Interview	S-E	VFI Factor 1	VFI Factor 2	VFI Factor 3
RTC	-5.56 (3.16) .001	-10.66 (3.48) <.001		-4.54 (4.10) .021
S-E		-5.10 (4.47) .017	6.08 (5.52) .022	
VFI Factor 1			11.18 (3.26) <.001	6.12 (4.29) .003
VFI Factor 2				-5.06 (3.92) .007

Table 12: Comparison of Part 2.1 scores for the final interview. Reported are the mean difference in measures across participants (standard deviation) and p-value.

The other potential interactions were not statistically significant. There was not a significant interaction of measure and gender, $F(2.159,23.754) = .642$, $p = .546$, $\eta^2 = .055$. In addition, there was a non-significant three way interaction (time, measure, gender), $F(4.024, 44.267) = .753$, $p = .562$, $\eta^2 = .064$. Finally, the interaction of time and gender was non-significant, $F(1.947,21.417) = 3.303$, $p = .057$, $\eta^2 = .231$. However, as with the interaction of time and measure, this nearly significant result warranted further investigation.

No statistically significant differences were found for women's scores between time points. However, some statistically significant differences were found for men's scores. Men's overall scores (across all measures) were lower in the initial interview (M= 7.121, SD=2.588) than in the midpoint (M= 9.293, SD= 1.864) and final interviews (M= 9.079, SD= 2.495), with

p= .011 and .010, respectively. Despite being non-significant, the women showed the same trend as the men, with lower scores in the initial interview (M= 8.475, SD=2.396) than the midpoint (M= 8.870, SD= 1.725) and final interviews (M= 8.724, SD= 2.310).

Qualitative Analysis

Trends by gender over time for RTC can be seen in Figure 6. This figure shows the average increase in RTC for both genders between the initial and midpoint interviews. From the midpoint to the final interview, men continue to show an increase (although smaller than initial to midpoint), and women show a decrease (although still higher than at the initial interview). For both genders, this indicates an effect of monitoring alone. However, the difference in genders between the midpoint and final interviews warrants further investigation. Perhaps the feedback is more supportive in increasing RTC for men, and ways to improve its support for women should be explored.
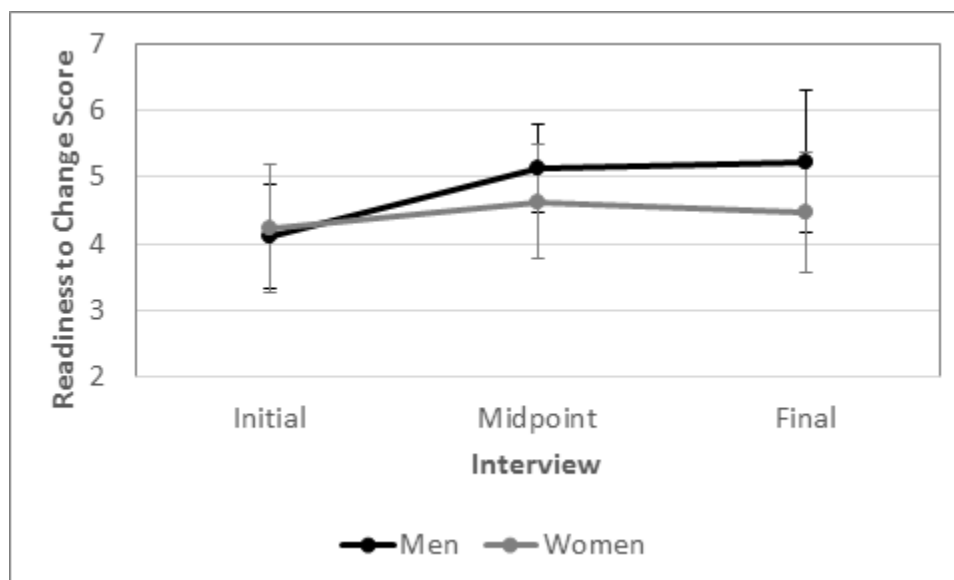
Figure 6: Average Readiness to Change scores over time by gender for Part 2.1.

Trends by gender over time for S-E can be seen in Figure 7. This figure shows an interesting difference between genders. Scores for men increase across the three interviews, while scores for women decrease across the three interviews. Like RTC, this suggests that potentially the current feedback is more supportive in increasing S-E for men, and ways to improve its support for women should be explored.
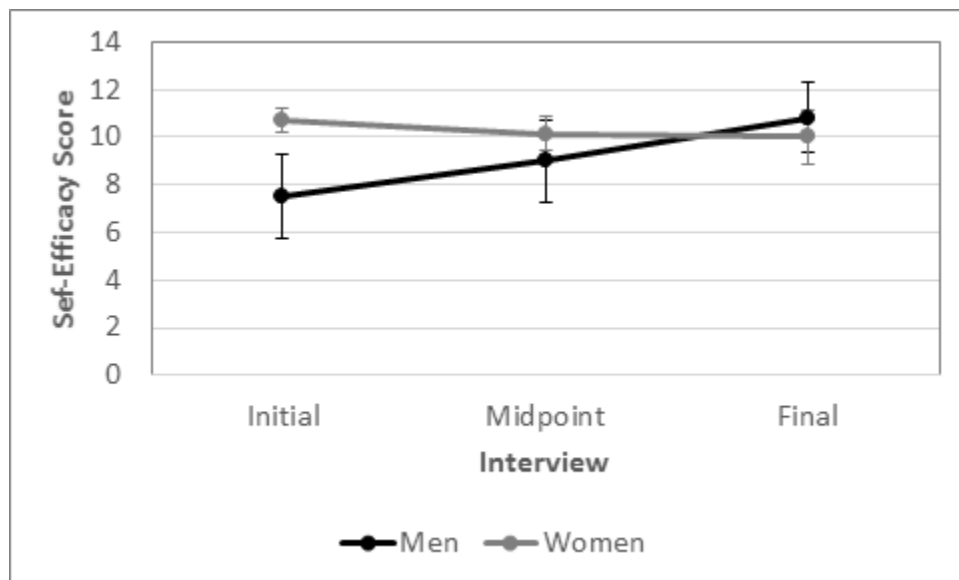

Figure 7: Average Self-Efficacy scores over time by gender for Part 2.1.

Trends by gender over time for VFI can be seen in Figures 8-10. For Factors 1 and 2, an increase in score indicates an increase in perceived vocal fatigue. Factor 1 is characterized by a feeling that the voice is "tired," which may lead to a reduction in further voice use, and Factor 2 is characterized by physical discomfort, such as a sore throat (Nanjundeswaran et al., 2015). Therefore, the average initial increase in both genders on both Factor 1 and Factor 2 may be consistent with increased awareness increasing the perception of vocal fatigue. The average decrease for men on both factors, and the average decrease for women on Factor 2 indicates that

feedback may have helped reduce some of this perceived fatigue. The continued average increase

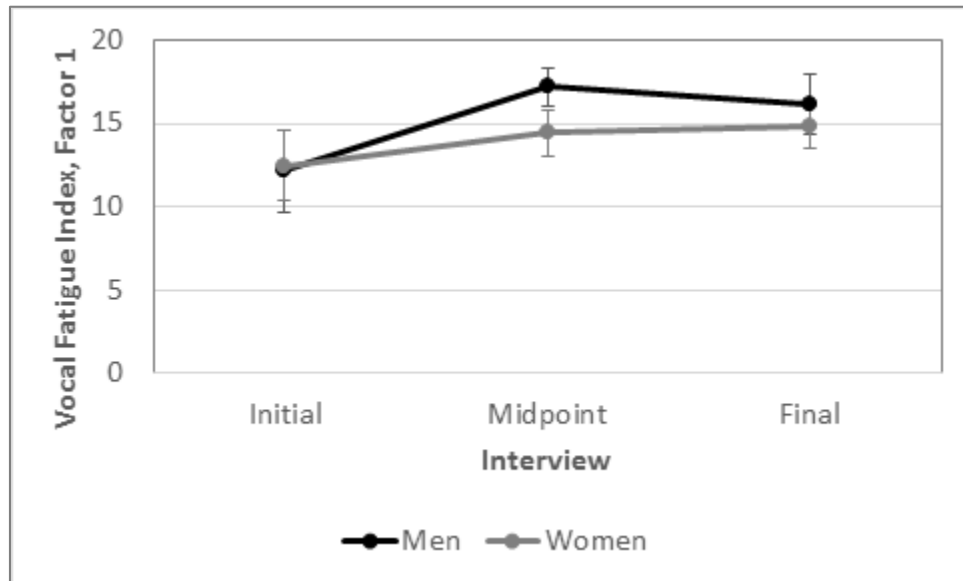for women on Factor 1 does not follow this trend.

Figure 8: Average Vocal Fatigue Index, Factor 1 scores over time by gender for Part 2.1.
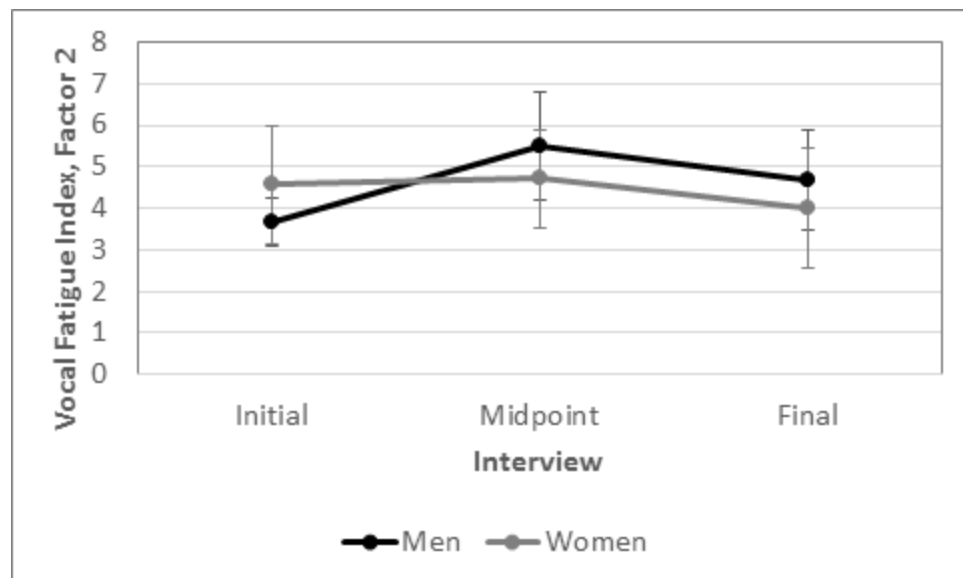

Figure 9: Average Vocal Fatigue Index, Factor 2 scores over time by gender for Part 2.1.

On the other hand, an increase in score indicates a <u>decrease</u> in perceived vocal fatigue for

Factor 3. Factor 3 is characterized alleviation of fatigue symptoms with vocal rest

(Nanjundeswaran et al., 2015). Therefore, men showed an average decrease in perceived vocal

fatigue from the initial to the midpoint interview, and an increase at the final interview (although

still less than at the initial interview). This overall increase in score indicates a reduction in vocal

fatigue. On the other hand, women showed no change (on average) between the initial and

midpoint interviews, and an increase in vocal fatigue at the final interview.

While not specifically explored in this study, the increasing vocal fatigue experienced by

women on Factors 1 and 3 may be due to the VLT used in Part 2.1. Some participants (such as

P2104) reported continuing to increase their vocal intensity over time in the study due to

increased comfort with the task, and this type of behavior may have led to increased perception

of fatigue within the study, which artificially influenced the questionnaire results. Therefore,

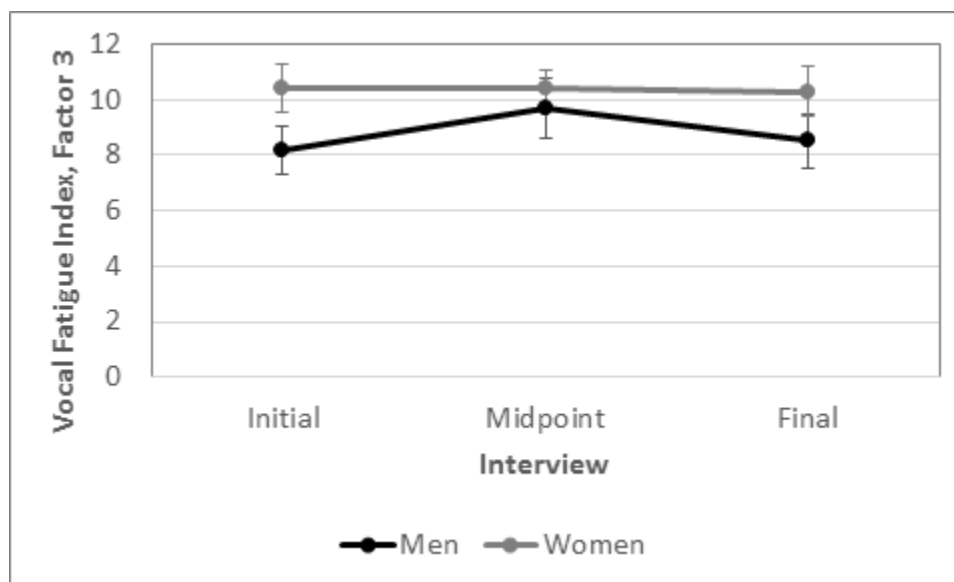further work in this area should attempt to tease these effects apart.

Figure 10: Average Vocal Fatigue Index, Factor 3 scores over time by gender for Part 2.1.

Part 2.2

The assumption for sphericity was met for all variables, so no corrections were needed. The results of the mixed ANOVA are reported below.

Main Effects

There was a significant main effect of measure, $F(4,12) = 55.464$, $p < .001$, $\eta^2 = .949$. Almost all measure pairings were statistically significantly different at or below $p= .043$ with a few exceptions. Statistically significant differences were found for the following comparisons: RTC (M= 6.514, SD= 2.023) and VFI Factor 1 (M= 22.467, SD= 5.330); RTC and VFI Factor 3 (M= 8.667, SD= 1.799); S-E (M= 10.133, SD= 3.226) and VFI Factor 1, VFI Factors 1 and 2 (M= 6.333, SD= 2.992), and VFI Factors 2 and 3. Again, this main effect was less meaningful because the scaling and underlying constructs measured by the questionnaires was different, and these differences probably contributed to the significant differences. The findings are consistent with the model of three factors of vocal fatigue because all three factors were statistically significantly different.

There was a non-significant effect of time, $F(2,6) = 1.374$, $p = .323$, $\eta^2 = .314$. However, due to the significance found in Part 2.1, further exploration for identifying trends was undertaken. The mean score across questionnaires was higher in the initial interview (M= 10.319, SD= 3.067) than the final interview (M= 9.830, SD= 4.388), but both were lower than the midpoint interview (M= 11.229, SD= 4.554). Relative trends are explored in more detail below. Corresponding individual results can be seen in Appendices P-R (Figures 143-145).

For RTC, this trend of increasing and then decreasing was seen in 4 of the 5 participants (80%). However, only 2 of the 5 participants maintained higher RTC in the final interview

compared with the initial interview (2 participants' scores were lower in the final interview). This difference in trend from Part 2.1 should be further explored in future studies. While there were only 5 participants in Part 2.2, these individuals are occupational voice users, and are therefore closer to the target user population.

For S-E, less of a trend was noted. Only 2 individuals (40%) showed this trend, and both of them ended with lower S-E in the final interview than the initial interview. One individual showed a steady increase over the three interviews, one showed a decrease, and one had a decrease at the midpoint interview and a return to baseline in the final interview. These inconsistent suggest that more work with occupational voice users in the real-world environment are needed to understand the relationship between PVM and S-E.

For VFI Factor 1, a total of 3 participants (60%) followed the trend of an increase for midpoint followed by decrease in the final interview. Two of these individuals continued to have higher scores in the final interview than the initial, indicating more vocal fatigue. This is consistent with the findings in Part 2.1. Taking the findings from Parts 2.1 and 2.2 together, it is possible that voice monitoring alone draws greater attention to vocal fatigue and therefore participants report greater fatigue. With feedback, they are able to reduce that fatigue by at least some margin, and future work should explore how to further reduce perceived fatigue.

For VFI Factor 2, a total of 3 participants (60%) followed a similar trend of an increase followed by a decrease in fatigue. However, these results showed greater promise. In total, 4 of the 5 participants (80%) demonstrated a decrease in Factor 2 from the initial to the final interview, indicating reduced vocal fatigue. This is a positive finding that with more participants can hopefully achieve statistical significance. For VFI Factor 3, a total of 2 participants (40%)

followed a similar trend. However, only one participant (20%) showed an increase in Factor 3 from initial to final interview (indicating less fatigue).

No significant difference was found for gender, $F(1,3) = 2.242$, $p = .231$, $\eta^2 =.428$. This finding indicates that differences found were not due to differences between genders.

Interactions

There was a non-significant interaction of time and measure, $F(8,24) = 1.513$, $p = .205$, $\eta^2 =.335$. Unlike Part 2.1, this result was not approaching statistical significance.

Unlike Part 2.1, there was a significant interaction of measure and gender, $F(4,12) = 3.610$, $p = .037$, $\eta^2 =.546$. Post hoc pairwise comparisons revealed the same trends for men and women, with men showing greater variation between measures. Women showed an average difference of -14.632 (SD= 2.296) between RTC and VFI Factor 1, $p= .003$. In addition, women showed an average difference of 14.111 (SD= 4.980) between VFI Factors 1 and 2, $p= .034$. Finally, women showed an average difference of 10.667 (SD= 1.269) between VFI Factors 2 and 3, $p= .001$.

On the other hand, men showed an average difference of -24.052 (SD= 2.296) between RTC and VFI Factor 1, $p= .001$. In addition, men showed an average difference of 19.167 (SD= 4.980) between VFI Factors 1 and 2, $p= .025$. Finally, men showed an average difference of 18.500 (SD= 1.269) between VFI Factors 2 and 3, $p< .001$.

These findings were significant, even with a small number of participants, but the same interaction was non-significant in Part 2.1. Therefore, further investigation with a larger number of participants is needed to determine the nature of the interaction between gender and measure.

In addition, the interaction of time and gender was non-significant, $F(2,6) = 2.142$, $p =$ .199, $\eta^2 = .417$. Finally, there was a non-significant three way interaction (time, measure, gender), $F(8,24) = 1.238$, $p = .320$, $\eta^2 = .292$.

Qualitative Analysis

Trends by gender over time for RTC can be seen in Figure 11. This figure shows the average increase in RTC for both genders between the initial and midpoint interviews. From the midpoint to the final interview, women continue to show an increase (although smaller than initial to midpoint), and men show a decrease (although still higher than at the initial interview). This trend is the opposite of Part 2.1, where women showed the decrease in RTC from the midpoint to the final interview. This finding suggests that men may need additional feedback support to continue to increase RTC. This difference in trends across the two phases of Part 2 suggests further exploration to uncover the true relationship is needed.
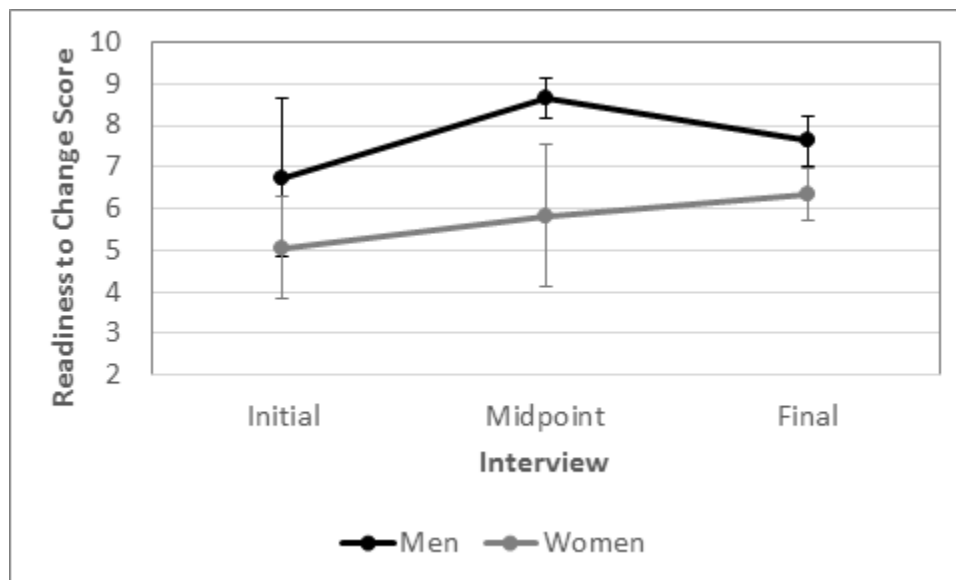
Figure 11: Average Readiness to Change scores over time by gender for Part 2.2.

Trends by gender over time for S-E can be seen in Figure 12. As with RTC, the findings with S-E are inconsistent with the findings from Part 2.1. Both genders show an average increase from the initial to the midpoint interview, and then an average decrease in the final interview. For both genders, the decrease in the final interview is below the average baseline value, indicating decreased S-E (opposite of the hypothesized change). This finding suggests that further exploration of the value of S-E as an outcome measure for PVM should be explored. Perhaps a more tailored self-efficacy scale for voice disorder prevention (rather than an adapted general S-E scale) would lead to different results.
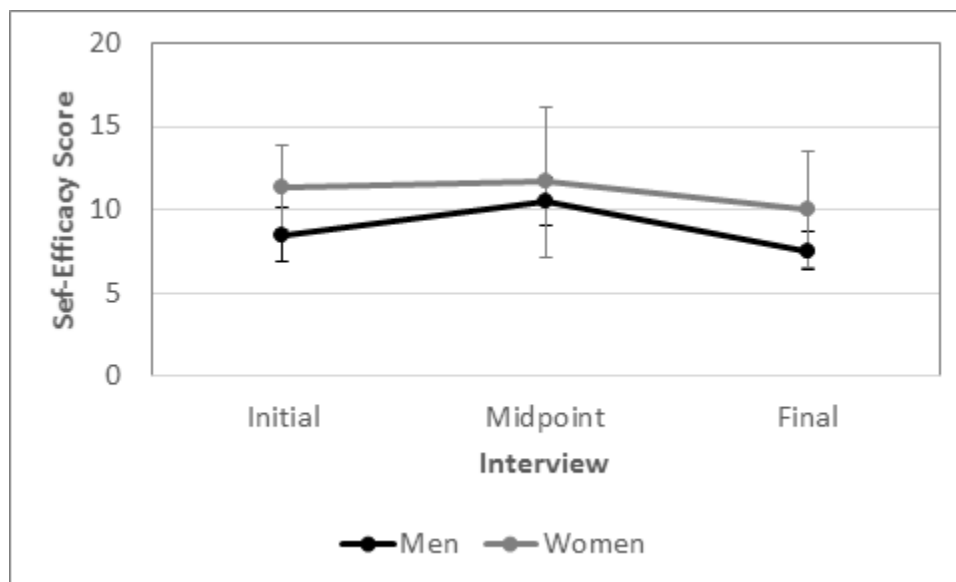
Figure 12: Average Self-Efficacy scores over time by gender for Part 2.2.

Trends by gender over time for VFI can be seen in Figures 13-15. For Factors 1 and 2, an increase in score indicates an increase in perceived vocal fatigue. Factor 1 is characterized by a feeling that the voice is "tired," which may lead to a reduction in further voice use, and Factor 2 is characterized by physical discomfort, such as a sore throat (Nanjundeswaran et al., 2015). Women had an average increase in both factors from the initial to midpoint interview, and men

had an average increase for Factor 1 (Factor 2 remained constant). As suggested in Part 2.1, the average increases in Factor 1 and Factor 2 may be consistent with increased awareness increasing the perception of vocal fatigue. As in Part 2.1, men showed average decreases on both factors from the midpoint to the final interview, indicating that feedback may have helped reduce some of this perceived fatigue. The continued average increase for women on both factors does not follow this trend and should be further explored in future studies.

On the other hand, an increase in score indicates a <u>de</u>crease in perceived vocal fatigue for Factor 3. Factor 3 is characterized alleviation of fatigue symptoms with vocal rest (Nanjundeswaran et al., 2015). Both genders demonstrated an average decrease from the initial to the final interview, suggesting an increase in this type of vocal fatigue. However, the gender difference at the midpoint interview shows potentially different responses to voice monitoring alone: women experienced an increase in perceived fatigue (decrease in score) and men experienced a decrease in fatigue (increase in score). However, both genders showed an overall increase in perceived Factor 3 fatigue, which again may be attributed to increased awareness.
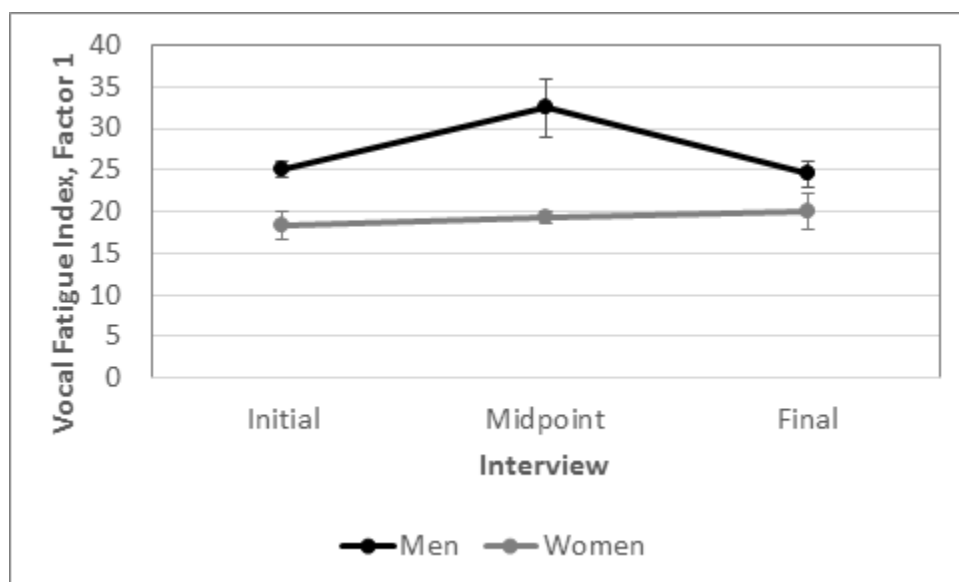
Figure 13: Average Vocal Fatigue Index, Factor 1 scores over time by gender for Part 2.2.
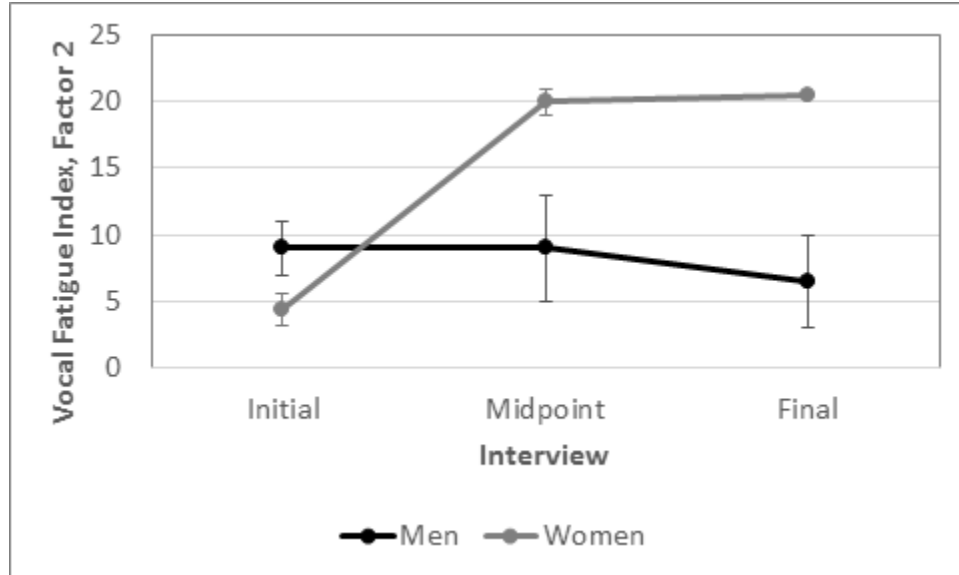
Figure 14: Average Vocal Fatigue Index, Factor 2 scores over time by gender for Part 2.1.
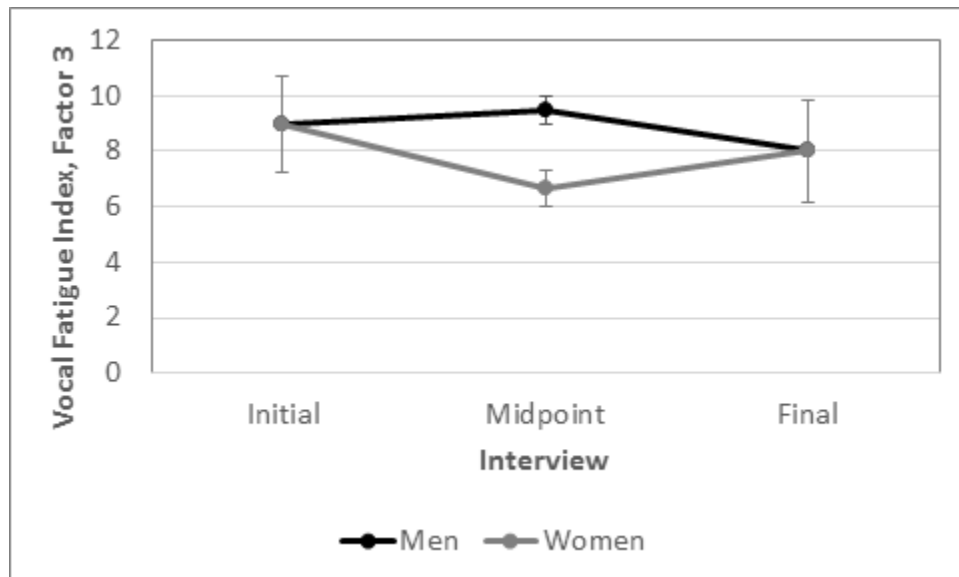
Figure 15: Average Vocal Fatigue Index, Factor 3 scores over time by gender for Part 2.1.

Summary

Extracting Design Requirements for Conveying Feedback

At the end of Part 1, vocal intensity (loudness), pauses, pitch strength (quality), and first moment specific loudness (strain) were identified as measures of interest by occupational voice users participating in the study. Rather than a single display for each measure, participants expressed that they preferred a layered structure which included a basic comparison between all days, a more detailed comparison between days, and finally a zoomed in version of each day. Feedback from participants in Part 2 suggested that strain might be the most helpful measure, with all other measures demonstrating some level of helpfulness. Participants liked that the displays were user-friendly and easy to read, but wanted clearer definitions of measures and suggestions on how to improve their voices based on the feedback measures. Interview results also indicated that occupational voice users are not a homogenous group, and feedback should be flexible in addressing a range of goals. Participants were more interested in relative trends rather than absolute values, especially because they did not have context to aid with delayed recall when comparing feedback across days. Additional features were suggested that should be explored in future iterations.

Identifying Changes in Voice Behavior Management after Receiving Feedback

The results from Part 2 indicate that PVM can be used to elicit behavior change in occupational voice users and future occupational voice users. Some of these changes were active, including trying to pause more when speaking and increasing/decreasing vocal intensity. Some of these changes manifested as changes in thinking. Many participants reported increased

awareness of the need to preserve the voice, especially from fatigue. Participants also shared insights into how they interpreted the feedback to better understand their own voices. Finally, some participants reported limited behavior change and/or little observed change in their feedback. Future research should attempt to determine what factors lead to greater vocal behavior change in some individuals than others, in an effort to encourage more occupational voice users to actively engage in PVM to reduce the risk of developing future voice disorders.

In addition to reported changes by participants during the semi-structured interviews, the questionnaire results suggested some behavior change as well. For Part 2.1, there was a main effect of time showing a general increase in scores on measures from the initial to the midpoint interview, but no statistically significant change in scores from the midpoint to the final interview. While not statistically significant in Part 2.2, a similar trend was found.

This is an interesting finding that possibly suggests greater behavior change from monitoring alone versus adding feedback. However, further examination suggests that some change does occur from the midpoint to the final interview that should not be discounted. For example, RTC increased from the initial to the final session for 9 of 13 participants (69.2%). Even though the increase was greater from initial to midpoint, there was still an overall increase in RTC following feedback. As another example, from Part 2.2, VFI Factor 2 scores increased at midpoint for 3 participants (60%), but when comparing initial versus final scores, 80% experienced a decrease. This decrease indicates less Factor 2 vocal fatigue (physical discomfort).

Both Part 2.1 and Part 2.2 showed a statistically significant main effect of measure. While this result is interesting, it does little to contribute to the overall purpose of the study. It was anticipated that these measures would measure different aspects of behavior change, and they all use different scales.

Finally, a significant interaction of measure and gender was found for Part 2.2. This finding indicated that men demonstrate greater differences in scores on the different measures, which may need to be taken into account in future studies.

Finally, the three-way interaction between time, measure, and gender was explored qualitatively as no statistically significant results were found. The trends suggest some measure-specific differences between the genders over time, some of which varied between Parts 2.1 and 2.2. This suggests that further exploration into this three-way interaction may be necessary to understand how to maximize the feedback to best support a wide variety of potential PVM users.

Conclusions

Four measures (loudness, strain, quality pauses) were chosen as feedback, presented on a series of three displays each, with each display displaying a different aspect of the measure. Based on feedback from participants, strain was the most helpful, but the other measures were helpful as well. Participants reported increased voice awareness and behavior change as part of the study. However, they felt that better measure descriptions, suggestions on how to improve voice use, and the ability to add context to recordings would enhance the user experience. Additionally, there was a common trend of increases in RTC, S-E, and VFI scores after recordings (both midpoint and final scores were generally higher than initial scores). While this is a positive direction for RTC and S-E scores, the increased VFI scores indicate greater perceived vocal fatigue. This may be attributed to increased awareness, but needs to be explored in greater detail in future studies.

CHAPTER 5: Quantification of Voice Changes After Feedback

Study Overview

The goal of this study was to determine whether Preventative Voice Monitoring (PVM) would impact vocal behavior in occupational voice users and future occupational voice users. Chapter 4 discussed the design of the feedback displays (Aim 1) and insight on the sensitivity of behavior change questionnaires to internal changes experienced by users of PVM (Aim 2). Chapter 5 discusses observed changes in voice production (Aim 3) as a result of PVM.

Part 2 Methods

Using the feedback displays developed in Part 1, the researcher tested the feedback display prototypes in Part 2. Potential target users completed vocal loading tasks (VLTs), tasks designed to induce vocal fatigue, to determine whether objective feedback influenced later voice production. Part 2 consisted of two phases: 1) laboratory testing (reading aloud for 15 minutes at an elevated vocal intensity) and 2) field testing (classroom lectures).

Participants

Fourteen participants enrolled in Part 2.1, but only 13 participants met the inclusion criteria of the study (6M, 7F; M=22.6 years, SD=9.15 years, Range=18-48 years). P2108 was excluded from the study because she did not meet the inclusion criteria. All remaining participants were either current occupational voice users or future occupational voice users.

An additional five participants enrolled in Part 2, Phase 2 (Part 2.2), and all participants met all the inclusion criteria (2M, 3F; M=42.6 years, SD=12.40 years, Range=33-62 years). All

participants were currently instructors at Michigan State University who taught courses at least two days per week. More detailed summary information is provided in Chapter 4: Part 2 Methods: Participants.

Procedures

Part 2 consisted of data collection over 11 individual sessions for each participant with no session lasting more than one hour. These sessions consisted of three interview sessions and eight recording sessions. Only the recording sessions will be discussed in Chapter 5. All recording sessions featured a vocal loading task (VLT). The VLT consisted of reading aloud from "Charlotte's Web" for 15 minutes in Part 2.1, and lecturing to a classroom in Part 2.2. For Part 2.1, participants were instructed to "Read aloud as though you are reading to a classroom and want to be heard by all the children," which are the same instructions given in a prior study to elicit louder speech (Hunter et al., 2015). Loud speech is one type of task that is common to induce vocal fatigue in participants (Solomon, 2008). This instruction allowed the participants in Part 2.1 to approximate the vocal effort of the course instructors, although for a shorter time. Further descriptions of the VLTs can be found in Chapter 4: Part 2 Methods: Recording Sessions.

Recording Sessions

All recording sessions followed the same general pattern. Before and after each VLT, the participant completed a series of short tasks to allow the researcher to assess changes pre- to post-VLT. The short tasks were completed in a randomized order across sessions. In one task, participants produced three sustained /ɑ/ vowels of approximately 3-5 seconds in duration at a

comfortable pitch and loudness. The average reading from a sound pressure level (SPL) meter (at a distance of one meter) for the third /ɑ/ vowel prior to reading was used as the calibration level for the dosimeter. Participants also read the Rainbow Passage (Fairbanks, 1960) aloud at a comfortable loudness, and rated vocal fatigue on a scale from 1 to 10 (1 = not at all; 10 = the most extreme). For Part 2.1, the short tasks were performed in the same sound-treated booth used for the VLTs. For Part 2.2, the short tasks before and after the VLT were performed in a location of the instructor's choosing (e.g., laboratory, office, or classroom).

All participants (Parts 2.1 and 2.2) received feedback based on measurements from the accelerometer signal. Because VLTs in Part 2.2 occurred in the classroom, where ambient noise would affect measures from the microphone recording, the researcher used the accelerometer signal for feedback analysis and presentation. This is consistent with prior dosimetry studies (Hunter & Titze, 2010). However, previous studies validated the use of pitch, pitch strength, and first moment specific loudness only for acoustic (audio) signals (e.g., Shrivastav & Camacho, 2010; Shrivastav et al., 2012). The focus of this study was to assess the feasibility of providing feedback on voice use to people performing a VLT, and therefore these measures were taken from the accelerometer for the feedback, but correlations of these measures from the two types of signals (accelerometer, audio) were conducted to determine the generalization of these measures.

The data for the feedback was analyzed using a combination of Goldwave (Goldwave Inc., St. John's, NL, Canada) and MATLAB (The MathWorks, Natick, MA, USA) processes. For the steps used in these analyses, see Appendix S (Part 2.1) and Appendix T (Part 2.2). Analysis was different for pitch strength and pauses between Part 2.1 and 2.2 due to constraints using MATLAB scripts in Part 2.2 with audio files greater than one hour in length. See below for basic descriptions of these analyses.

First Moment Specific Loudness

The researcher ran this analysis in a consistent manner across both Part 2.1 and 2.2. In summary, the researcher extracted the sustained vowels before and after the VLT. Next, the researcher input the middle 500ms of each vowel into a MATLAB script to estimate spectral moments using custom MATLAB scripts (Kopf et al., 2013).

Pitch Strength: Part 2.1

Pitch strength analysis was conducted using Auditory-SWIPE' (Camacho, 2007) for all participants. Because this is a time-intensive analysis in MATLAB, the researcher completed this analysis for the entire VLT for participants in Part 2.1, but only for the initial and final 15-minute lecture segments for Part 2.2. Participants saw the average pitch strength reported minute-by-minute so they could see the variability and change over time, if any.

Pitch Strength: Part 2.2

Due to the time-intensive nature of running Auditory-SWIPE', time constraints did not allow pitch strength to be calculated for full class periods (due to data collection being twice a week). Rather, participants were given feedback on the first and last 15 minute segments from each class period so they could see how/if pitch strength varied from the beginning to the end.

Pauses: Part 2.1

Pauses were calculated using modifications of the same MATLAB script between Part 2.1 and 2.2. For Part 2.1, pauses were calculated based on the pitch strength and pitch output. If

the pitch of the output segment was <80 Hz, it was determined to be a segment of silence. The researcher chose this cutoff value because many of the error values given were due to noise at 77 Hz. Because the output of pitch strength is in 0.01 second segments, 100+ consecutive silence segments were considered to be pauses of 1 second or greater and were reported in the feedback. The researcher chose this cutoff value based on the existing literature. In turn taking, gaps longer than a second are rare (Heldner & Edlund, 2010; Wilson & Wilson, 2005). In addition, silences less than a second are considered to be at the word/phrase boundary level (Titze et al., 2007), whereas silences of a second or more are considered to be at a sentence level. This is also above the threshold for long pauses in both healthy controls (502ms) and speakers with ataxic dysarthria (767ms) (Rosen et al., 2010). Therefore, these pauses of longer duration (one second or greater) might be associated with effort to pause on the part of the speaker.

Pauses: Part 2.2

For Part 2.2, because pitch strength analyses could not be conducted on the entire classroom recording due to time constraints, pauses were based on dB analyses. Those segments that were <80 Hz and below a given dB threshold (varied across participants) were considered to be segments of silence. Because the output of the dB analysis was in 0.1 second segments, 10+ consecutive silence segments were considered to be pauses of 1 second or greater and were reported in the feedback.

dB Level: Part 2.2

The average dB level was shown for every three seconds of recording for ease of displaying results. For the feedback displays, the average dB value replaced periods of silence.

Analysis

Aim 3

Aim 3: To quantify changes in the voice after receiving feedback.

For statistical analysis, a multi-measure MATLAB script, custom designed in this laboratory for a prior study (Schloneger & Hunter, 2016; Schloneger, 2014), was used to analyze phonation time, dB level, fundamental frequency, pitch, and pitch strength from both the accelerometer and audio recordings. First moment specific loudness was evaluated using the same MATLAB script from the feedback analysis. Analyses were completed separately for Part 2.1 and Part 2.2.

In addition to addressing the three Aim 3 hypotheses, correlational comparisons of pitch, pitch strength, and first moment specific loudness from microphone (audio) and accelerometer signals were conducted. These comparisons assessed the appropriateness of using these measures from an accelerometer signal.

Hypothesis 1

Hypothesis 1 states: It was hypothesized that occupational voice users would improve voice production in response to feedback, and these improvements would manifest as decreases in one or more of the following: vocal intensity, voicing time, and/or F0.

To address Hypothesis 1, the researcher completed a mixed ANOVA. There were two within-subjects factors: time (baseline, feedback) and measure (vocal intensity, voicing time, fundamental frequency). These measures were taken from the accelerometer signal. In addition,

gender was the between-subjects factor. Because the focus was on whether there was a change between baseline and feedback recordings, the researcher compared the 15-minute VLT average of the three baseline sessions with the 15-minute VLT average of the five feedback sessions.

Hypothesis 2

Hypothesis 2 states: It was hypothesized that increasing vocal fatigue would result in the following changes in objective voice quality measures: increasing strain (increasing first moment specific loudness) and/or increasing breathiness (decreasing pitch strength).

To address Hypothesis 2, the researcher completed a stepwise linear regression, with an entry value of 0.05 and removal value of 0.10. Pitch strength and first moment specific loudness were the independent variables, and participants' self-rated vocal fatigue was the dependent variable. The researcher took the measures from the audio signal. Pitch strength and first moment specific loudness were averaged across the three sustained /ɑ/ vowels for each time point (e.g., before the VLT in the first baseline recording session). These average values and their corresponding vocal fatigue scores were entered into the analysis as separate data points. Therefore, each participant had a maximum of 16 data points (eight before VLTs, eight after VLTs), with some missing data due to loss of recordings.

Hypothesis 3

Hypothesis 3 states: Changes in breathiness and vocal strain pre- to post-vocal loading task would be greater for baseline tasks than tasks with feedback.

To address Hypothesis 3, the researcher completed a mixed ANOVA. In this analysis, there were two within-subjects factors: time (baseline, feedback) and measure (pitch strength,

first moment specific loudness). The researcher took the measures from the audio signal. In addition, gender was the between-subjects factor. Because the focus was on whether there was a change between baseline and feedback, the researcher first averaged each measure (pitch strength, first moment specific loudness) from the three /ɑ/ vowels before each VLT, and did the same for the three /ɑ/ vowels after each VLT. Then, the researcher took the difference (before minus after) of the average values for a given measure in a given session, resulting in one average difference score for pitch strength and one for first moment specific loudness per session. Then, the researcher compared the average difference score of the three baseline sessions with the average difference score of the five feedback sessions.

Results and Discussion

Correlation Results

One of the important things explored in this study was whether one can reliably measure pitch (as measured using Auditory-SWIPE'), pitch strength, and first moment specific loudness from a reduced bandwidth (accelerometer) signal.

Comparison of pitch from accelerometer and audio signals

The correlation of pitch from the two channels of the recording, averaged across all tasks (vowels, rainbow passage, reading) and all speakers, was found to be 0.901, suggesting that this measure can be used for accelerometer signals – at least for perceptually normal adult voices.

Comparison of fundamental frequency and pitch

The correlation of fundamental frequency and pitch, averaged across all tasks (vowels, rainbow passage, reading) and all speakers, was found to be 0.961 for the accelerometer signal and 0.898 for the audio signal. These high correlations suggest that both measures are measuring the same phenomenon. However, this correlation should be treated with caution as this does not hold true for voice signals with reduced voice quality (Shrivastav, Eddins, & Kopf, 2014).

Comparison of pitch strength from accelerometer and audio signals

The correlation of pitch strength from the two channels of the recording, averaged across all tasks (vowels, rainbow passage, reading) and speakers, was found to be 0.683, suggesting that this measure is not appropriate for use with accelerometer signals. For two of the three initial participants, P2101 and P2102, the correlations were at or above 0.79, indicating a stronger similarity between the two analyses, but this did not to hold when more participants' data was included in the analysis. It is widely recognized that pitch strength can vary with spectral shape (e.g., Fastl & Zwicker, 2007); thus, a difference in pitch strength values when it is estimated from the lower bandwidth accelerometer signal vs. a higher bandwidth microphone signal is not surprising.

Comparison of first moment specific loudness from accelerometer and audio signals

The correlation of first moment specific loudness from the two channels of the recording, averaged across all tasks (vowels, rainbow passage, reading) and speakers, was found to be 0.399, suggesting that this measure is not appropriate for use with accelerometer signals. Again, correlations were higher with P2101 (0.54) and P2102 (0.41), but not high enough to justify the

use of this measure to estimate strain from accelerometer signals in future studies. Since spectral

moments are directly related to the spectral bandwidth, these findings are of no surprise.

Aim 3 Results and Discussion

Data from the recording sessions was used in the analyses below. While most participants

did not experience any issues with the recordings, some issues led to loss of data for other

participants. Data loss in voice monitoring studies is not an uncommon problem (Hunter & Titze,

2010; Schloneger & Hunter, 2016).

In Part 2.1, P2106 and P2109 experienced a technical issue that caused loss of part of a

recording from one session. P2111 experienced infrequent difficulties with the recording

equipment, ultimately resulting in nearly complete data loss for one session and partial data loss

for another.

In Part 2.2, P2201 was sick during the last two recording days but still chose to record on

those days. This data was omitted from Hypothesis 1 and 3 analyses to avoid her reduced voice

quality from negating any positive changes that might be captured in the other three feedback

days. However, data from these days was included in Hypothesis 2, correlations between voice

measures and vocal fatigue ratings, because these recordings were of suboptimal voice and

provided addition variation in analysis. On the final feedback recording day, there was an issue

with recorder set-up that led to P2202's data being completely lost. However, this participant

was willing to come back for a 9th recording session (approved by the MSU HRPP), and this data

was used in place of his day 8 data. Issues on days 6 and 8 of recording led to data loss for

P2203. P2204 had the most data loss of any participant- 4 of the 8 recording days' data (2

baseline recordings, 2 feedback recordings) was unusable or only partially usable. Finally, due to a scheduling issue with P2205's courses, only seven recordings were made.

Even though there were missing data points for individual sessions, this did not affect the overall analysis for Hypothesis 1 and Hypothesis 3. Missing data was treated as missing at random, and if a data point from a session was missing, the average for a recording type included all non-missing data. For example, for P2204, "baseline sessions" consisted of one data point (one complete, two missing). While this was not an ideal way to do the analysis, it allowed for comparisons across participants (although some caution should be used in interpretation).

Hypothesis 1

The assumption for sphericity was met for all variables for Part 2.1 and 2.2 analyses. Therefore, no corrections were needed. The results of the mixed ANOVA are reported below. Due to the high correlations between F0 and pitch for accelerometer signals, pitch was used in place of F0 for this analysis.

Part 2.1 Main Effects

Participant specific results can be seen in Appendix U, V and W, for phonation time, vocal intensity, and pitch, respectively (Figures 146-148). The results of the mixed ANOVA found a significant main effect of measure, $F(2,22) = 815.148$, $p < .001$, $\eta^2 = 0.987$. Pairwise comparisons demonstrated that all three measures were statistically significantly different at $p < .001$. Phonation time (M= 78.97, SD= 3.641) was significantly higher than dB level (M= 60.845, SD= 5.033) and both measures were significantly higher than pitch in semitones (M= 41.952, SD= 5.510). However, as with the comparison of the questionnaire scores, this comparison adds

little to the overall question of the study. These measures would be expected to have different means because they are quantifying distinct attributes of the signal.

There was no significant main effect of time, $F(1,11) = 3.508$, $p = .088$, $\eta^2 = .242$. In addition, there was a non-significant effect of gender, $F(1,11) = .761$, $p = .402$, $\eta^2 =.065$.

Part 2.1 Interactions

There was a significant interaction of measure and gender, $F(2,22) = 34.677$, $p < .001$, $\eta^2 = .759$. Post hoc testing with univariate ANOVAs (all results for each measure collapsed across time) found that pitch was significantly lower for men (M= 36.598, SD= 2.162) than women (M= 46.618, SD= 2.159), $F(1,24) = 139.059$, $p < .001$, $\eta^2 =.853$, and even with alpha correction for multiple comparisons (0.05/3 = 0.017), this finding is still significant. This finding is consistent with prior literature that women speak at a higher pitch (fundamental frequency) than men (Titze, 2000).

Statistical analysis also found that dB level was significantly higher for men (M= 62.952, SD= 4.749) than women (M= 59.093, SD= 4.748), $F(1,24) = 4.265$, $p = .050$, $\eta^2 =.151$, but with alpha correction for multiple comparisons (0.05/3 = 0.017), this finding was non-significant. However, this finding of increased dB level is somewhat different from the literature, where a prior dosimetry study found that female teachers had a higher average dB level than male teachers (Titze & Hunter, 2015). However, the tasks were different, with the prior study in an actual workplace environment whereas this task involved artificially increasing dB level.

Finally, statistical analysis found that phonation time was not significantly higher for men (M= 80.462, SD= 3.471) than women (M= 77.658, SD= 3.472), although this finding approached significance, $F(1,24) = 4.215$, $p = .051$, $\eta^2 =.149$. The phonation times found in this

study were elevated from those reported in the literature, where average speaking levels are less than 30% and even choral singing is not greater than 40% (e.g., (Schloneger, 2014; Szabo Portela, Hammarberg, & Södersten, 2013). Further investigation into this discrepancy is warranted.

There were no significant interactions of time and gender, $F(1,11) = .067$, $p = .800$, $\eta^2 = .006$, or time and measure, $F(2,22) = 1.200$, $p = .320$, $\eta^2 = .098$. Finally, there was a non-significant three way interaction (time, measure, gender), $F(2,22) = 1.196$, $p = .321$, $\eta^2 = .098$.

Qualitative Analyses

In addition to the statistical analysis, qualitative analysis was completed to see where trends may be occurring that warrant further investigation in future studies with a larger number of participants. The following figures show the average results by session and by recording time (baseline, feedback) by gender with standard error bars to indicate the variability within genders.

Phonation time

Average phonation time (in percent total time) can be seen by session (Figure 16) and by recording type (Figure 17). Again, these values are elevated from prior literature, and absolute values should be interpreted with caution. However, the trends (Figure 16) demonstrate an initial increasing in phonation time across the three baseline sessions and a lesser degree of change during the feedback recordings. While the hypothesized direction for phonation time was a decrease with feedback, Figure 17 demonstrates a small decrease for men, but an increase for women. Future work should look into what cues may be given through feedback to encourage reduction of phonation time.
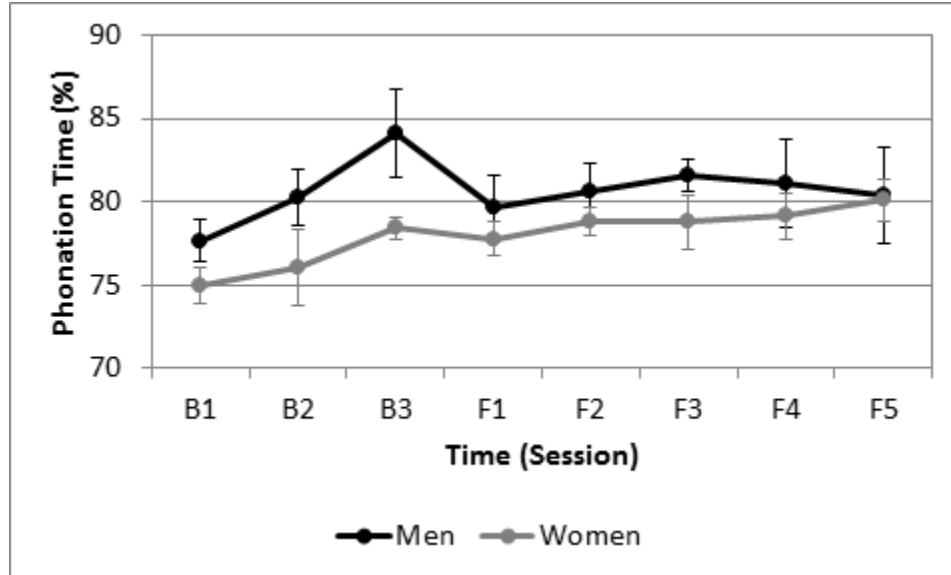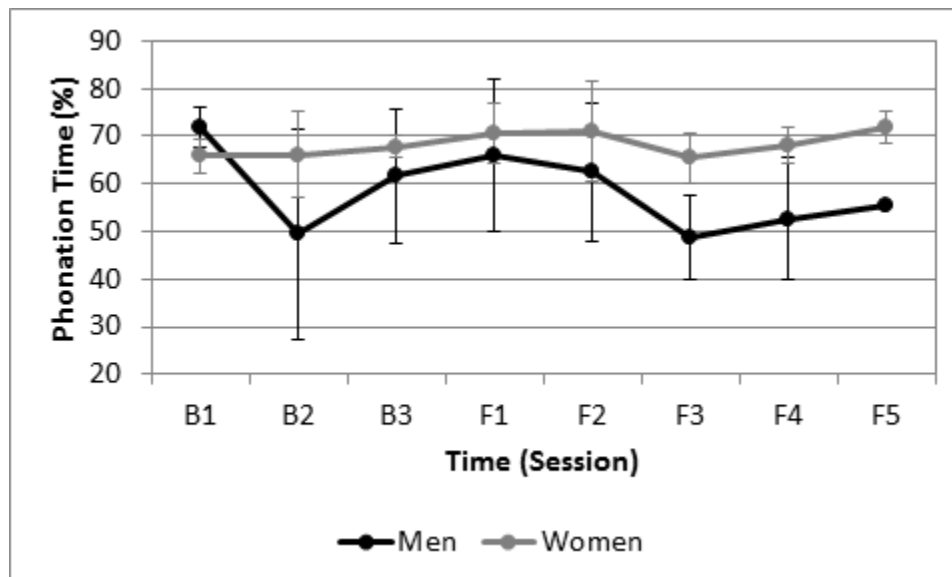
Figure 16: Average phonation time for each session by gender for Part 2.1.
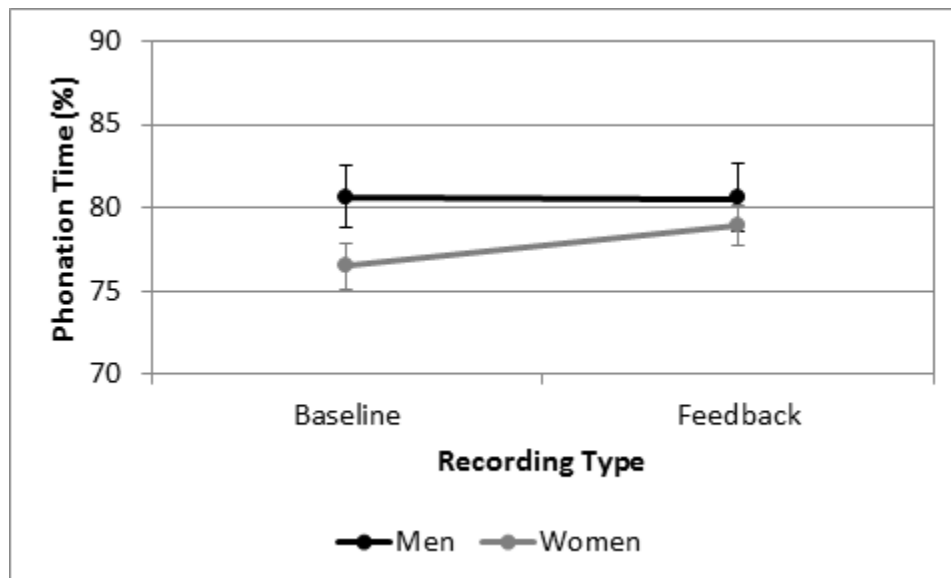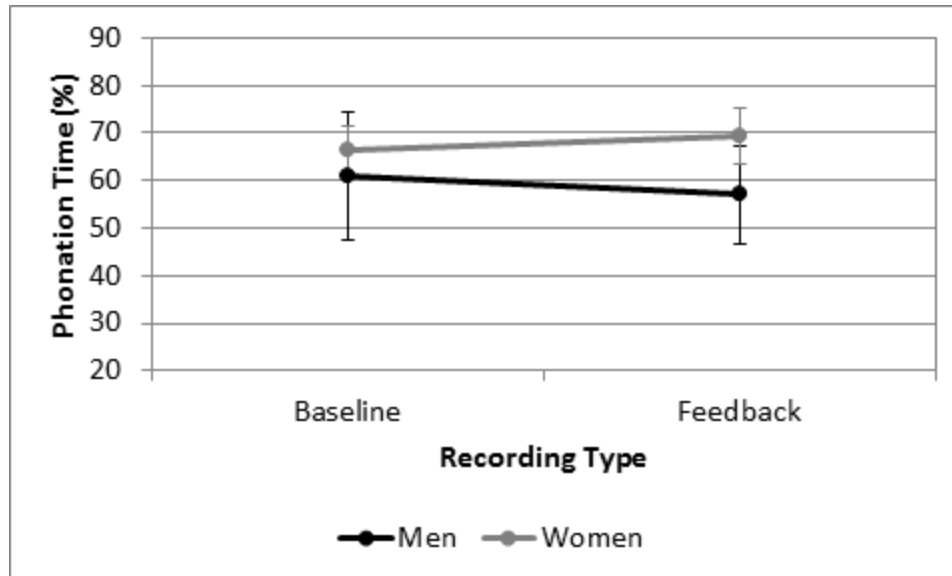

Figure 17: Average phonation time by recording type for Part 2.1. Results are reported separately by gender.

Vocal Intensity

Average vocal intensity can be seen by session (Figure 18) and by recording type (Figure 19). For baseline sessions, (Figure 18) both genders showed similar average intensity for the first and third sessions, but a deviation from that average in the second session (with an increase for men and a decrease for women). While the hypothesized direction for vocal intensity was a decrease with feedback, Figure 19 demonstrates an increase for both men and women. Prior results discussed in Chapter 4: Part 2 Results and Discussion highlighted the fact that some participants reported being "quiet talkers", and so this increase in dB level was seen as a positive trend for these speakers. This again highlights the need for further investigation of occupational voice user goals, which should be accounted for when assessing outcome measures.
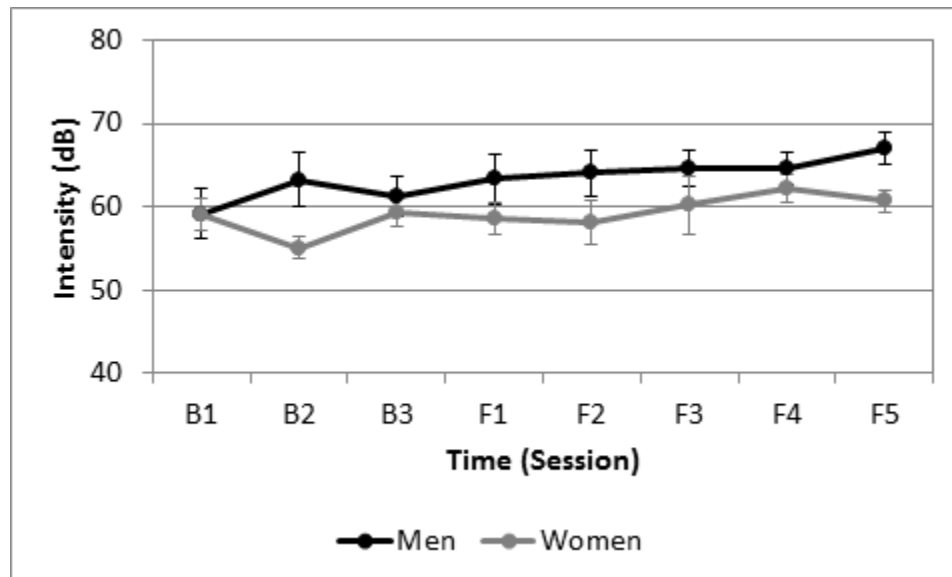
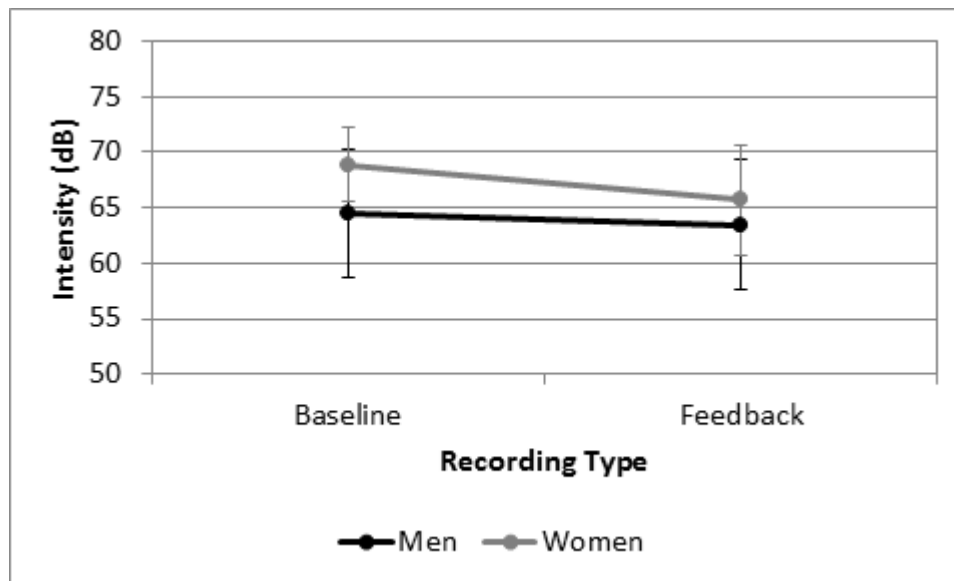Figure 18: Average vocal intensity for each session by gender for Part 2.1.
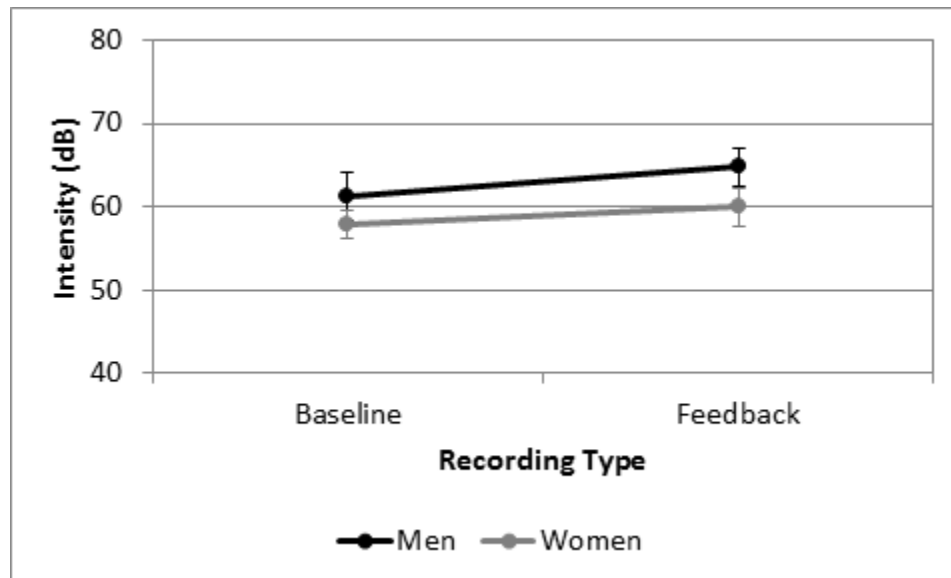
Figure 19: Average vocal intensity by recording type for Part 2.1. Results are reported separately by gender.

Pitch

Average pitch (in semitones) can be seen by session (Figure 20) and by recording type (Figure 21). The trends (Figure 20) demonstrate an initial decrease in pitch across the three baseline sessions for men, with women showing a more stable pitch across sessions. While the hypothesized direction for pitch was a decrease with feedback, Figure 21 demonstrates a small decrease for men (from a mean of 35.48 to a mean of 35.33), but a small increase for women (from a mean of 45.80 to a mean of 45.96). One interesting point was that this was one outcome measure that participants did not receive feedback on. While an overall decrease in pitch (fundamental frequency) can reduce one's risk of voice disorders, there is evidence that pitch changes with vocal fatigue, but there is controversy about the direction of pitch change (Solomon, 2008). Therefore, future work should explore whether pitch is appropriate feedback to give for PVM, and if so, how it should be presented to occupational voice users.
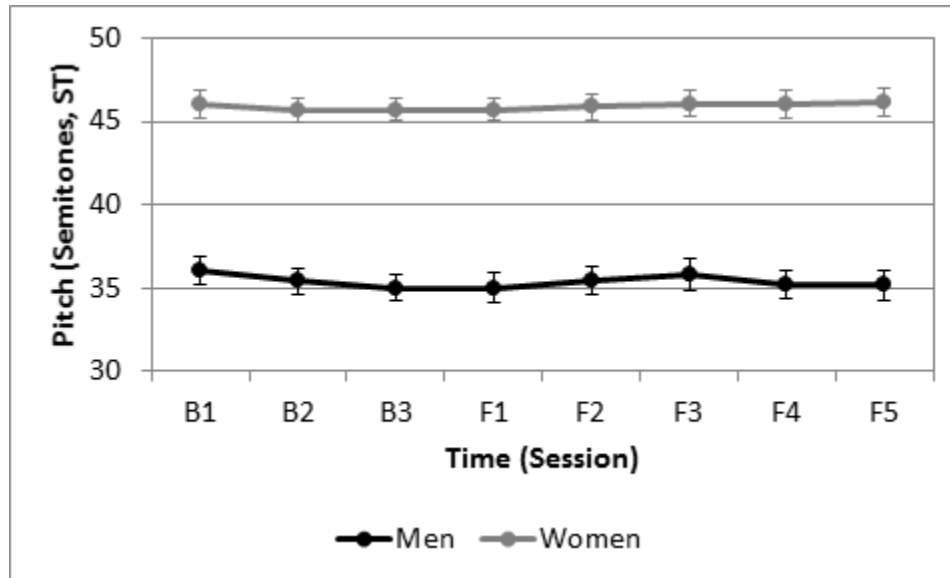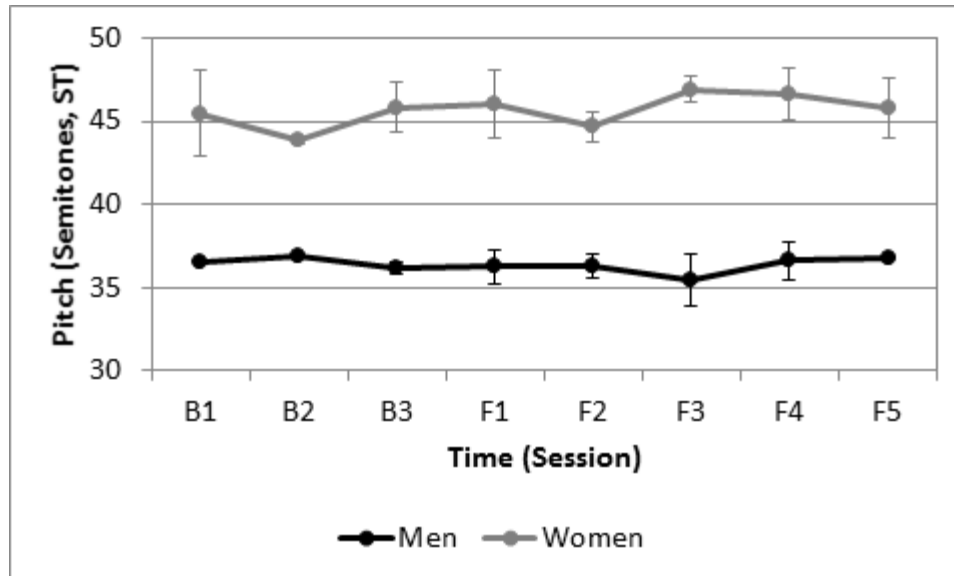
Figure 20: Average pitch for each session by gender for Part 2.1.
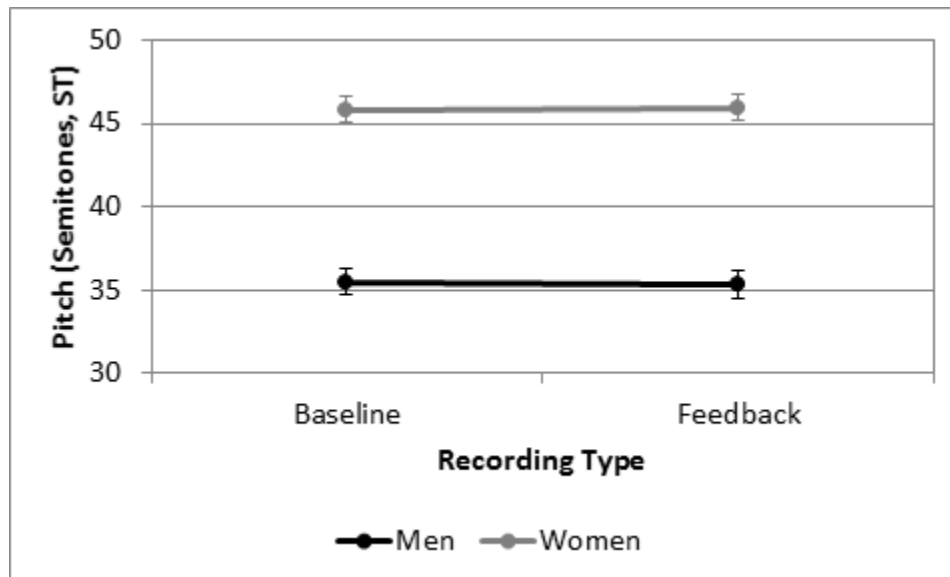
Figure 21: Average pitch by recording type for Part 2.1. Results are reported separately by gender.

Part 2.2 Main Effects

Participant specific results can be seen in Appendix X, Y and Z, for phonation time, vocal intensity, and pitch, respectively (Figures 149-151). The results of the mixed ANOVA found a significant main effect of measure, $F(2,6) = 20.116$, $p = .002$, $\eta^2 = .870$. The significant

difference was found between dB level (M= 65.000, SD= 4.385) and pitch (M= 41.000, SD= 4.847). Phonation time (M= 63.421, SD= 7.823) was not significantly different from either other measure. Again, this finding, though significant, did not provide additional insight because these differences were expected between measures.

There was a non-significant main effect of time, $F(1,3) = .123$, $p = .749$, $\eta^2 = .039$, gender was also non-significant, $F(1,3) = 2.199$, $p = .235$, $\eta^2 = .423$.

Part 2.2 Interactions

No statistically significant interactions were found. Measure and gender, significant in Part 2.1, was not significant, $F(2,6) = .769$, $p = .504$, $\eta^2 = .204$. In addition, there were no significant interactions of time and gender, $F(1,3) = .093$, $p = .781$, $\eta^2 = .030$, or time and measure, $F(2,6) = .527$, $p = .616$, $\eta^2 = .149$. Finally, there was a non-significant three way interaction (time, measure, gender), $F(2,6) = 2.767$, $p = .141$, $\eta^2 = .480$.

Qualitative Analysis

In addition to the statistical analysis, qualitative analysis was completed to see where trends may be occurring that warrant further investigation in future studies with a larger number of participants. The following figures show the average results by session and by recording time (baseline, feedback) by gender with standard error bars to indicate the variability within genders. Please note that standard error bars are not present for session 8 for the men. This is due to only one male instructor being recorded for the eighth session, so no error bars were generated.

Phonation time

Average phonation time (in percent total time) can be seen by session (Figure 22) and by recording type (Figure 23). Again, these values are elevated from prior literature, and absolute values should be interpreted with caution. Compared with Part 2.1, the trends (Figure 22) demonstrate greater variation session by session, with greater variation among men than women. Greater variation was expected, given the difference in nature between the VLTs in Part 2.1 and Part 2.2. While the hypothesized direction for phonation time was a decrease with feedback, Figure 23 demonstrates a small decrease for men, but an increase for women, the same as in Part 2.1. However, the results from Part 2.2 should be interpreted with greater caution. While a small segment (10 minutes) of the lecture was chosen where the speaker was mostly talking, these results may be influenced by segment selection, and factors outside the experimenter's control, such as the lecture content and disruptions during the lecture.

Figure 22: Average phonation time for each session by gender for Part 2.2.

Figure 23: Average phonation time by recording type for Part 2.2. Results are reported separately by gender.

Vocal Intensity

Average vocal intensity can be seen by session (Figure 24) and by recording type (Figure 25). Overall, both genders experienced a decrease in vocal intensity over the course of the study (Figure 25), but there was less of a trend for the men across the eight sessions (Figure 24).
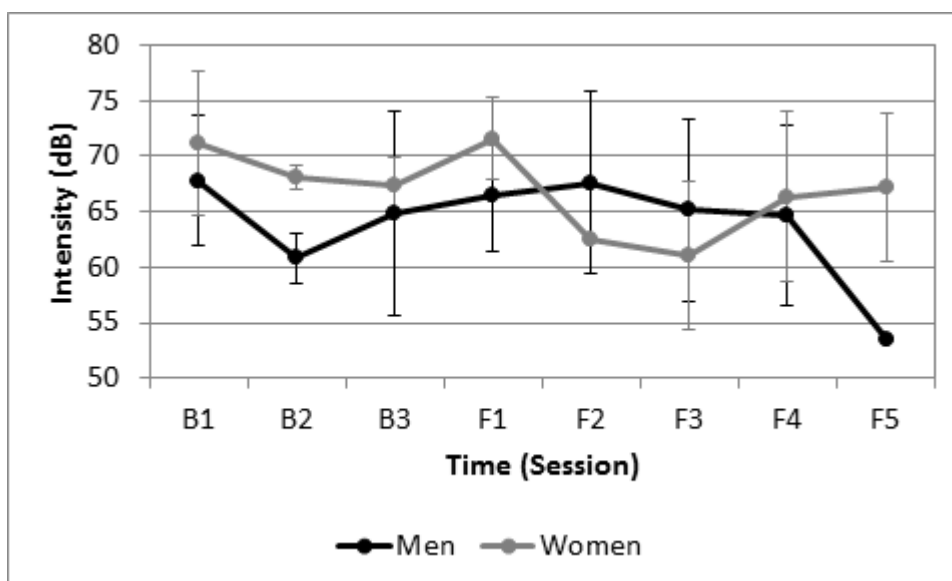
Figure 24: Average vocal intensity for each session by gender for Part 2.2.

This trend is in the hypothesized direction for reducing the risk of voice disorders. This measure should be monitored with a larger participant pool to see if the trend becomes statistically significant.

Figure 25: Average vocal intensity by recording type for Part 2.2. Results are reported separately by gender.

Pitch

Average pitch (in semitones) can be seen by session (Figure 26) and by recording type (Figure 27). In contrast with Part 2.1, the trends (Figure 26) show greater variability across session for women than men. While the hypothesized direction for pitch was a decrease with feedback, Figure 27 demonstrates similar trends to Part 2.1: a small decrease for men (from a mean of 36.53 to a mean of 36.26), but an increase for women (from a mean of 45.02 to a mean of 46.00). Therefore, future work should explore this measure with a greater number of participants to see if the trends become statistically significant. In addition, future work should explore whether pitch is appropriate feedback to give for PVM, and if so, how it should be presented to occupational voice users.

Figure 26: Average pitch for each session by gender for Part 2.2.
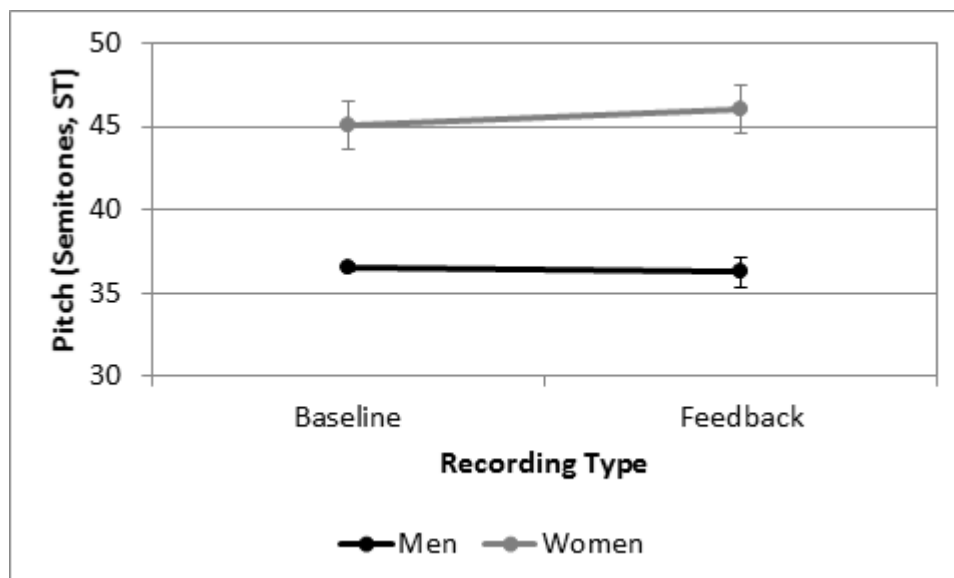
Figure 27: Average pitch by recording type for Part 2.2. Results are reported separately by gender.

Hypothesis 2

Part 2.1 Results

Stepwise linear regression with all participants did not yield a result (no variables were entered into the equation). However, because P2101 never reported a change in fatigue (rating was always 1), this participant was removed from the analysis, and it was run again with 12 participants' data. The result of Pearson correlations indicate that pitch strength and fatigue are correlated at r= 0.166, while first moment specific loudness is correlated with fatigue at r= 0.161. The results of the regression indicated that a single predictor, pitch strength, accounted for 2.8% of the variance (R^2 = 0.028, $F(1,185)$ = 5.261, p = .023). It was found that decreasing pitch strength significantly predicted increasing speaker fatigue (β = -4.182, p = .023).

Qualitative Analysis

In addition to the statistical analysis, qualitative analysis was completed to see where trends may be occurring that warrant further investigation in future studies with a larger number of participants. The following figures show the average results by session and by recording time (baseline, feedback) by gender with standard error bars to indicate the variability within genders.

The summary results can be seen in Figures 28-31. Figures 28 and 29 show separate linear regressions for pitch strength and first moment specific loudness, respectively, including P2101. Figures 30 and 31 show separate linear regressions for pitch strength and first moment, respectively, excluding P2101. It is of interest to note that there is a trend in the relationship of pitch strength and fatigue that is consistent with the hypothesis (increasing fatigue will be related with decreasing pitch strength). However, the trend for first moment specific loudness was not in

the hypothesized direction, because increasing first moment specific loudness is correlated with increasing vocal strain, but as fatigue increased, this measure decreased.

In all cases, there is still a great amount of unexplained variance in scores. This is not unlike prior findings for rating scales in voice quality work. Kreiman, Gerratt, Kempster, Erman, & Berke, (1993) described the high amount of variance in voice quality ratings, especially for those that were not at either end of the rating scale. There are a number of identified factors that play a role in this variation, including random errors, such as attention to task, and criterion (systematic) errors, such as how one uses the scale (Shrivastav, Sapienza, & Nandur, 2005). Shrivastav et al. (2005) found that these errors can be corrected by standardizing the ratings and averaging multiple ratings. To better understand the relationships between pitch strength and fatigue and first moment specific loudness and fatigue, an experimental design that allows for multiple ratings of the voice at the same moment, and later averaging and standardization in the analysis might be better able to determine the true relationships between these measures.
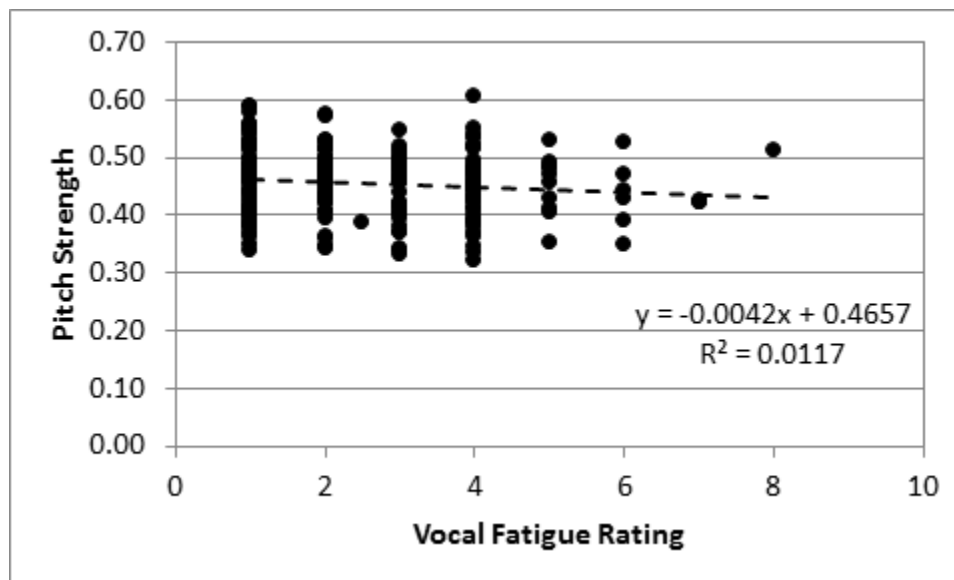
Figure 28: Linear regression of vocal fatigue rating and pitch strength including P2101.
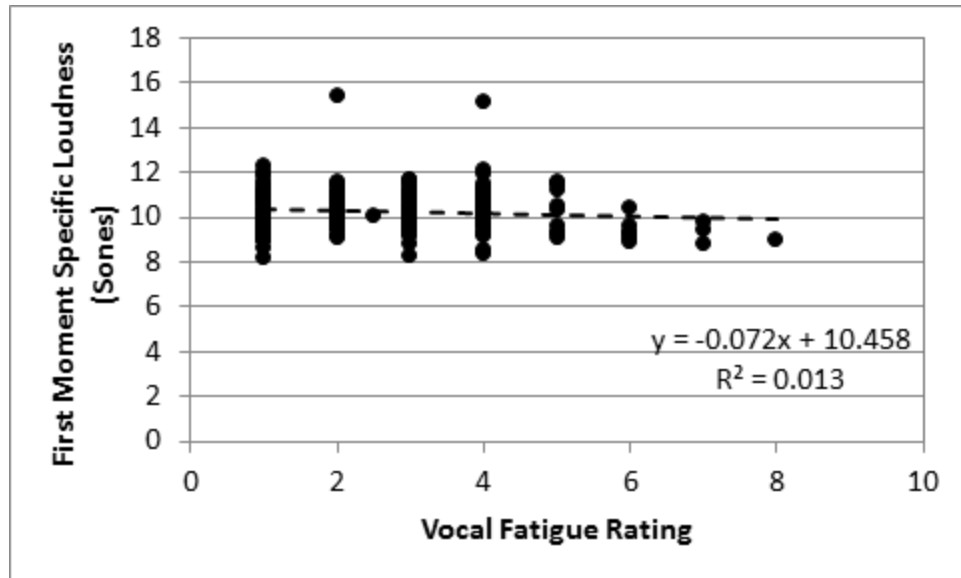
Figure 29: Linear regression of vocal fatigue rating and first moment specific loudness including P2101.
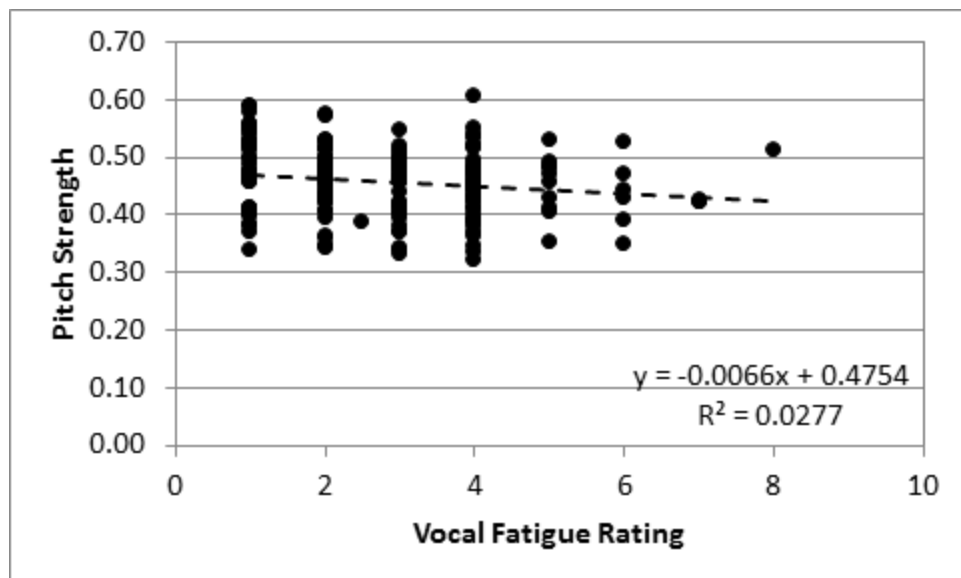
Figure 30: Linear regression of vocal fatigue rating and pitch strength excluding P2101.
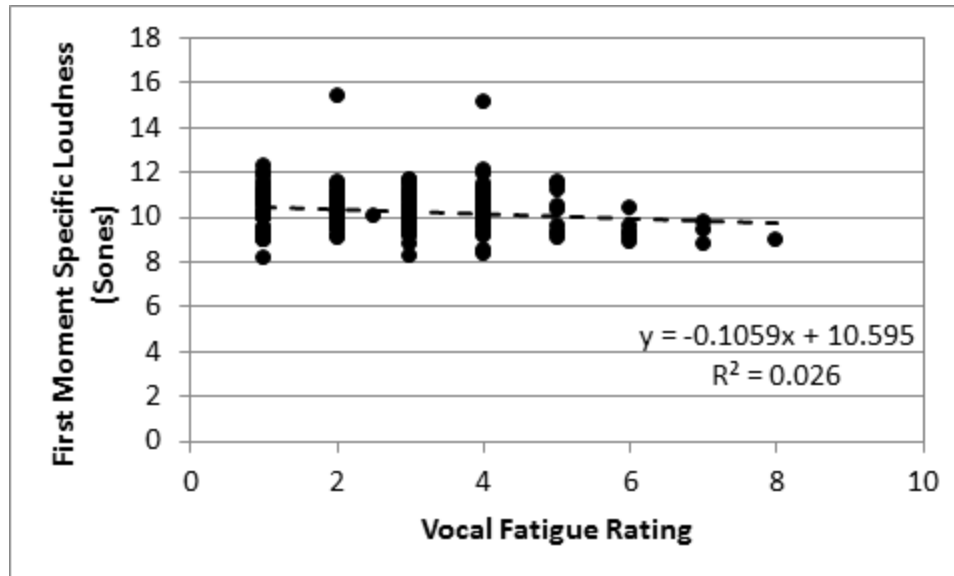
Figure 31: Linear regression of vocal fatigue rating and first moment specific loudness excluding P2101.

Part 2.2 Results

Stepwise linear regression with all participants did not yield a result (no variables were entered into the equation). However, the result of Pearson correlations indicate that pitch strength and fatigue are correlated at $r= 0.019$, while first moment specific loudness is correlated with fatigue at $r= 0.188$. The lack of entry into the regression may be due to a smaller number of data points than in Part 2.1 (where pitch strength was entered into the regression at $r= 0.166$), and therefore these measures should be re-examined with a greater number of data points in future research.

Qualitative Analysis

In addition to the statistical analysis, qualitative analysis was completed to see where trends may be occurring that warrant further investigation in future studies with a larger number of participants. The following figures show the average results by session and by recording time (baseline, feedback) by gender with standard error bars to indicate the variability within genders.

Please note that standard error bars are not present for session 8 for the men. This is due to only one male instructor being recorded for the eighth session, so no error bars were generated.

The summary results can be seen in Figures 32 and 33. Figures 32 and 33 show separate linear regressions for pitch strength and first moment specific loudness, respectively. It is of interest to note that the trend for the relationship of pitch strength and fatigue is the opposite of that in Part 2.1, and is therefore in the opposite direction of the hypothesis. First moment specific loudness, on the other hand, shows almost no trend and is relatively consistent across all vocal fatigue ratings.

Again, there is still a great amount of unexplained variance in scores. This is consistent with prior literature for subjective ratings. An experimental design with repeated ratings that can later be standardized in analysis would likely give a more consistent and clear picture of the actual relationships of pitch strength and first moment specific loudness with vocal fatigue.
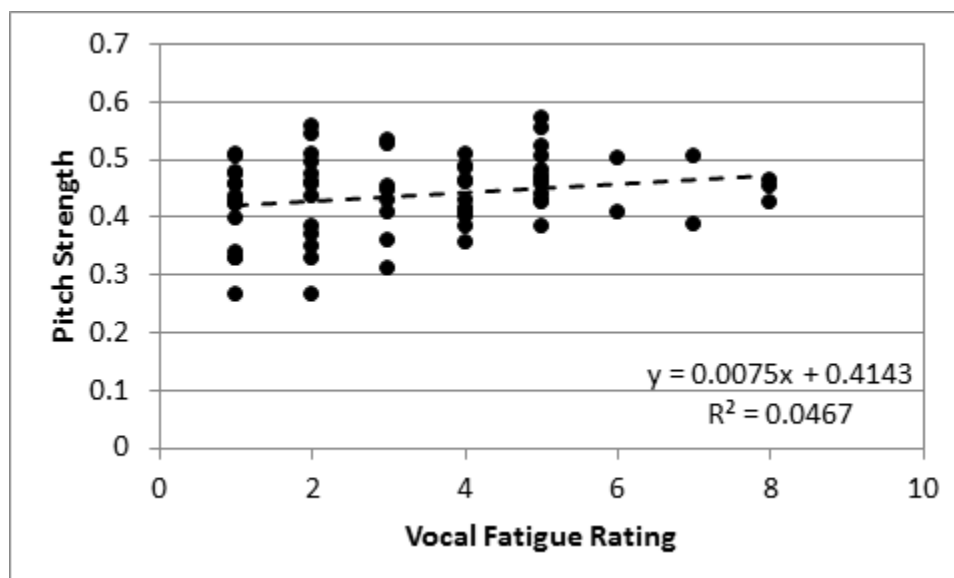
Figure 32: Linear regression of vocal fatigue rating and pitch strength.
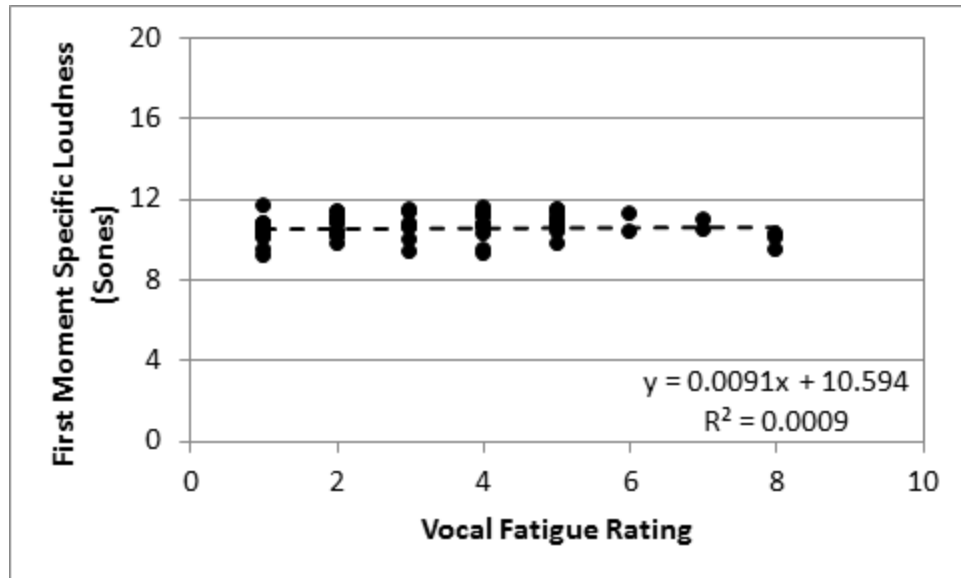
Figure 33: Linear regression of vocal fatigue rating and first moment specific loudness.

Hypothesis 3

Part 2.1 Main Effects

Participant specific results can be seen in Appendices AA and AB (Figures 152 & 153), respectively. No significant main effects were found. The results of the mixed ANOVA found no significant main effect of time, $F(1,11) = .415$, $p = .532$, $\eta^2 = .036$, and no significant main effect of measure, $F(1,11) = .759$, $p = .402$, $\eta^2 = .065$. There was no between subjects effect of gender, $F(1,11) = .093$, $p = .766$, $\eta^2 = .008$.

Part 2.1 Interactions

No significant interactions were found. There was a non-significant interaction of time and measure, $F(1,11) = .298$, $p = .596$, $\eta^2 = .026$. There was also no significant interaction of measure and gender, $F(1,11) = .141$, $p = .714$, $\eta^2 = .013$, or time and gender, $F(1,11) = .328$, $p =$

.578, η^2 =.029. Finally, there was no significant three-way interaction (time, measure, gender), $F(1,11) = .590$, $p = .459$, η^2 =.051.

Qualitative Analysis

In addition to the statistical analysis, qualitative analysis was completed to see where trends may be occurring that warrant further investigation in future studies with a larger number of participants. The following figures show the average results by session and by recording time (baseline, feedback) by gender with standard error bars to indicate the variability within genders.

Pitch Strength

The summary results can be seen in Figures 34 and 35. Average pitch strength can be seen by session (Figure 34) and by recording type (Figure 35). It was hypothesized that change scores (difference between the average pitch strength values from sustained /ɑ/ vowels) would decrease as a result of the feedback. This trend was seen for men (Figure 35) by the absolute value in change 0.02 to 0.01. The actual change was in the opposite direction between recording types: in baseline recordings, men had an increase in pitch strength post-VLT. This was not anticipated, but may be related to a warm-up effect (Vilkman et al., 1999). However, during feedback recordings, the hypothesized direction was seen- a decrease in pitch strength from pre- to post-VLT.

On the other hand, women displayed the opposite trend of the men. First, the absolute value of their change became greater from baseline to feedback, 0.02 to 0.03. In addition, during the feedback recordings, they showed the expected trend of a decrease in pitch strength following the VLT. During the feedback recordings, the opposite occurred. It is possible that

women were able to use this task as more of a warm-up in these later sessions, however, further
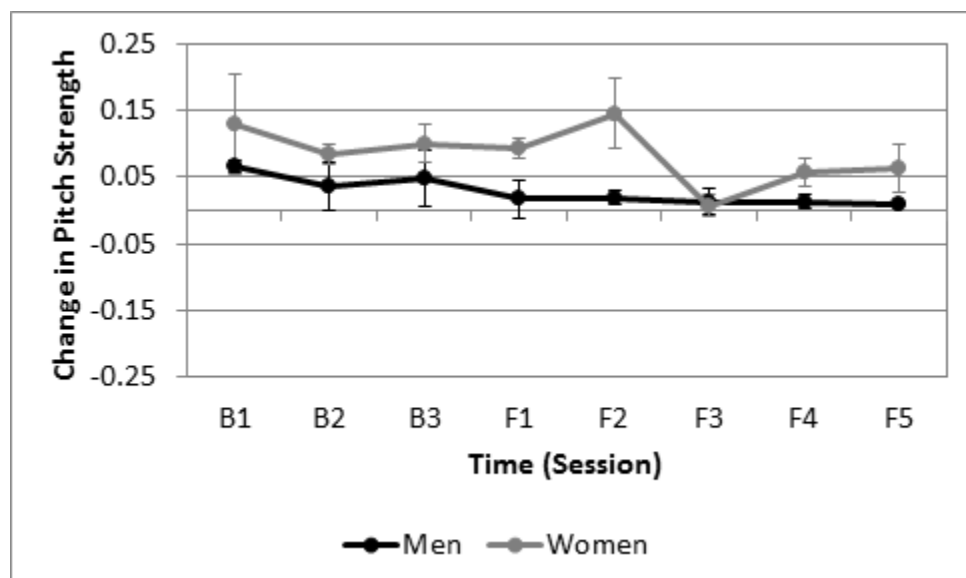
study should be done.
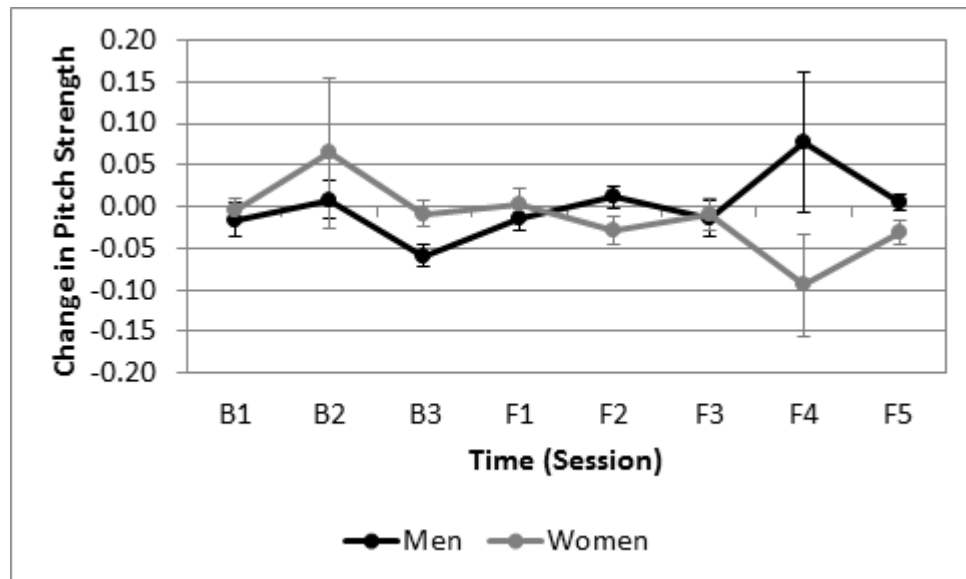
Figure 34: Average change scores (pre – post) for pitch strength by session in Part 2.1. Positive values indicate greater pitch strength before the vocal loading task.
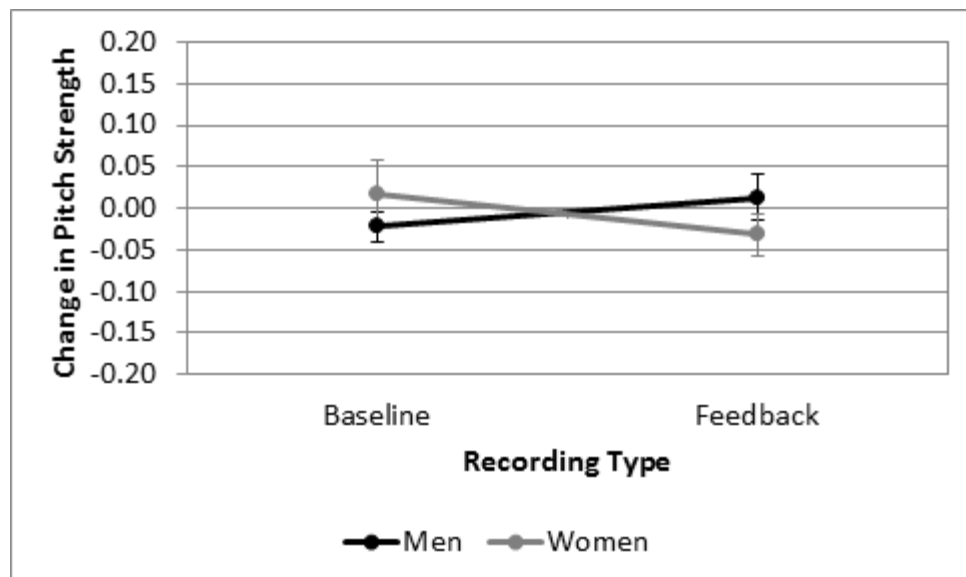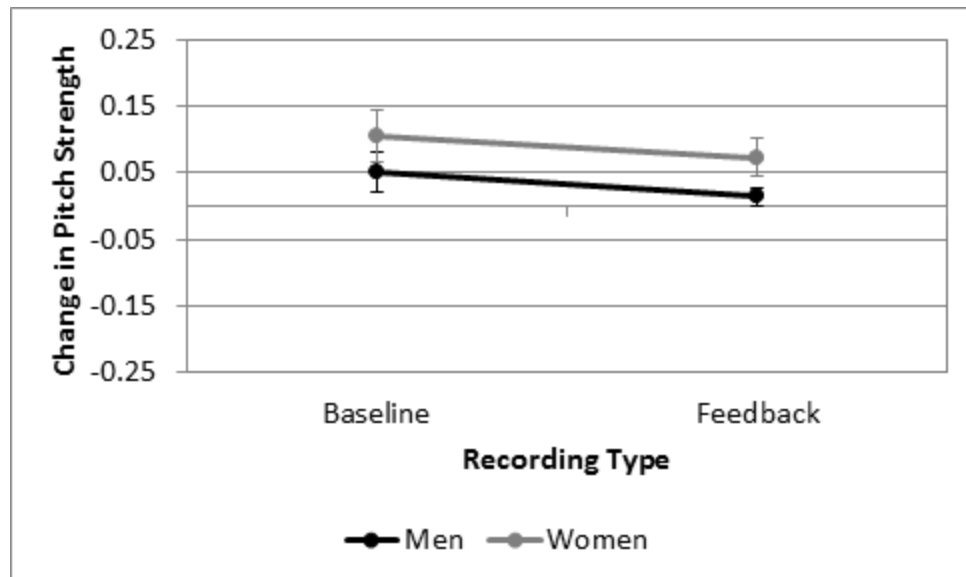

Figure 35: Average change scores (pre – post) for pitch strength by recording type in Part 2.1. Positive values indicate greater pitch strength before the vocal loading task.

Overall, these findings suggest some small change as a result of the feedback, but the direction of the results is mixed between the genders. Further study with a greater number of participants may be able to determine whether this effect holds for a more generalizable population.

First Moment Specific Loudness

The summary results can be seen in Figures 36 and 37. Average first moment specific loudness can be seen by session (Figure 36) and by recording type (Figure 37). It was hypothesized that change scores (difference between the average first moment specific loudness values from sustained /ɑ/ vowels) would decrease as a result of the feedback. The trends were the opposite of pitch strength. This trend was seen for women (Figure 36) by the absolute value in change 0.54 to 0.16. The actual change was in the opposite direction between recording types: in baseline recordings, women had a decrease in first moment specific loudness post-VLT (indicating less strain post-VLT). However, during feedback recordings, the opposite direction was seen- an increase in first moment specific loudness from pre- to post-VLT indicated by the negative value.

This finding is inconsistent with the pitch strength findings for women, if both measures are indicators of vocal fatigue. At baseline, they showed decreasing pitch strength and decreasing first moment specific loudness from pre- to post-VLT, potentially indicating increasing and decreasing vocal fatigue, respectively.

On the other hand, men displayed the opposite trend. First, the absolute value of their change became greater from baseline to feedback, 0.02 to 0.34. While this was not in the hypothesized direction of decreasing the change pre- to post-VLT, this finding suggests a greater

decrease in first moment specific loudness (strain) after feedback. This would be consistent with

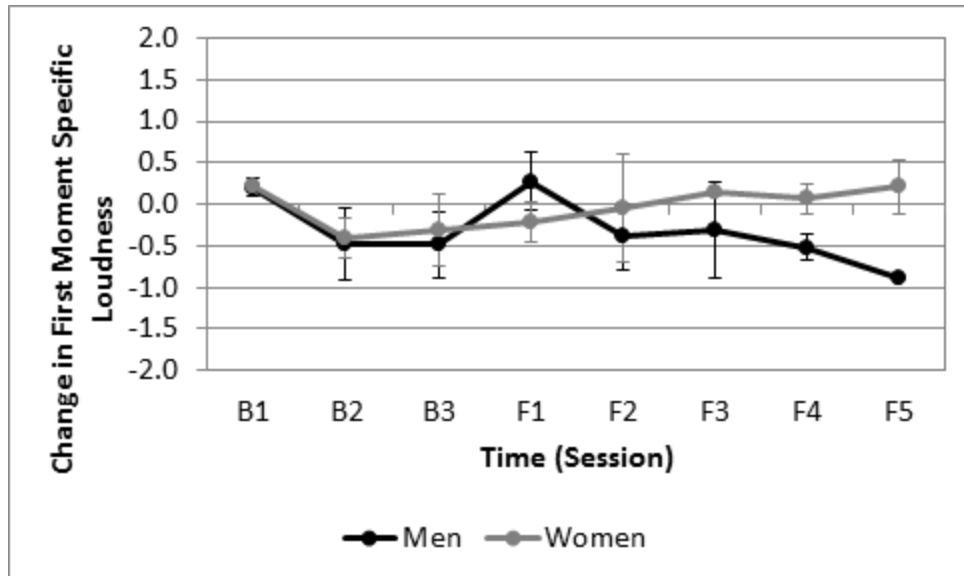a positive change as a result of the feedback.
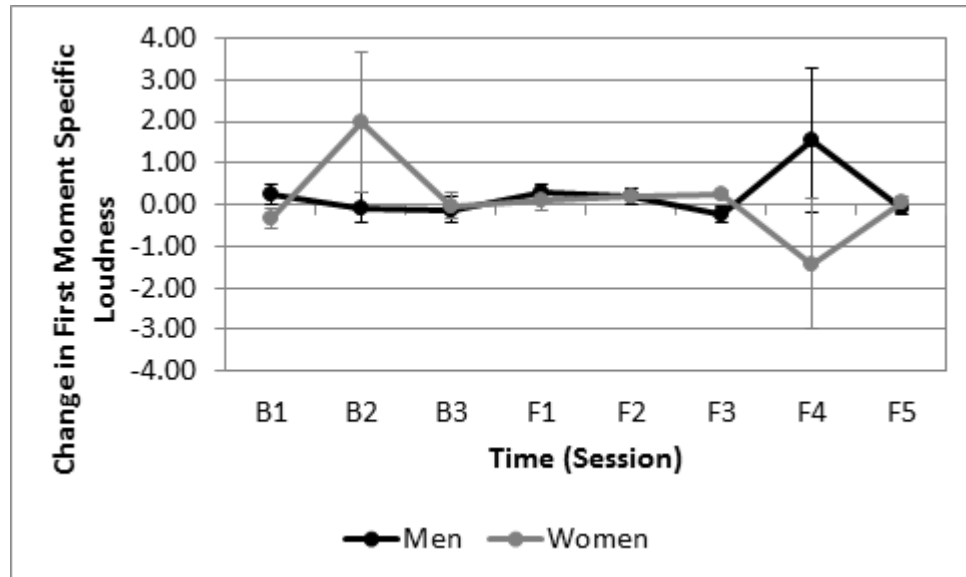
Figure 36: Average change scores (pre – post) for first moment specific loudness by session in Part 2.1. Positive values indicate greater first moment specific loudness before the vocal loading task.
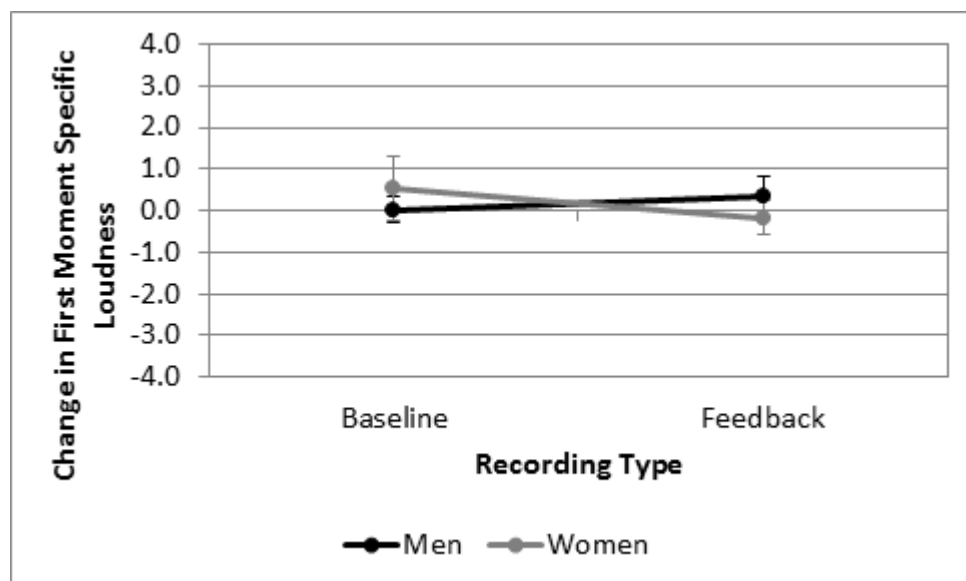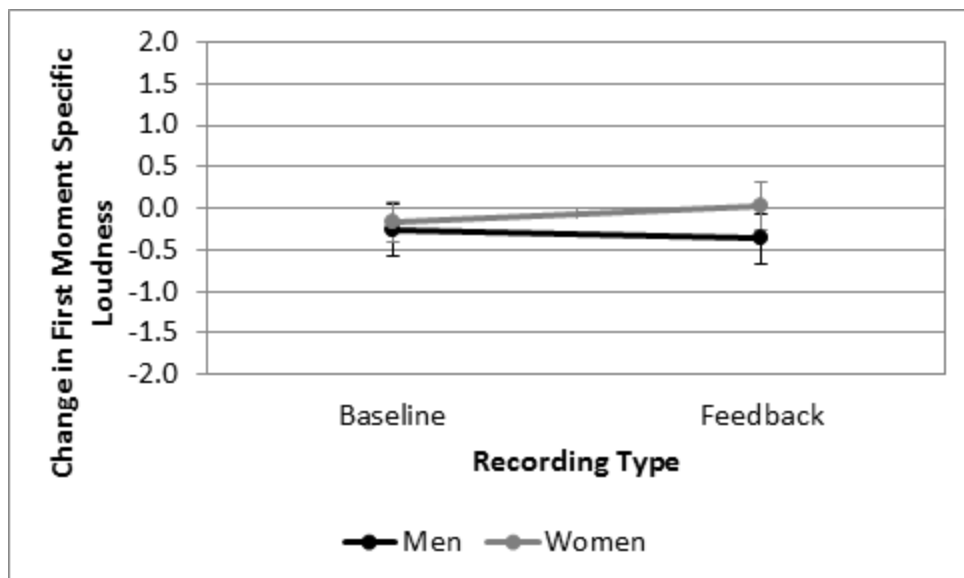

Figure 37: Average change scores (pre – post) for first moment specific loudness by recording type in Part 2.1. Positive values indicate greater first moment specific loudness before the vocal loading task.

Overall, these findings suggest some small changes as a result of the feedback, but the direction of the results is mixed between the genders. However, due to the small, non-statistically significant trends, this hypothesis needs to be substantiated with additional testing with a larger number of participants to determine if one or both measures are indicators of vocal fatigue (possibly associated with different fatigue factors).

Part 2.2 Main Effects

Participant specific results can be seen in Appendices AC and AD (Figures 154 & 155), respectively. As in Part 2.1, there were no statistically significant main effects. The results of the mixed ANOVA found no significant main effect of time, $F(1,3) = .014$, $p = .912$, $\eta^2 = .005$, and no significant main effect of measure, $F(1,3) = 5.392$, $p = .103$, $\eta^2 = .643$. There was no between subjects effect of gender, $F(1,3) = 1.201$, $p = .353$, $\eta^2 = .286$.

Part 2.2 Interactions

As in Part 2.1, there were no statistically significant interactions. The results of the mixed ANOVA found no interaction of time and measure, $F(1,3) = .515$, $p = .525$, $\eta2 = .147$, no significant interaction of measure and gender, $F(1,3) = .397$, $p = .573$, $\eta2 = .117$, and no interaction of time and gender, $F(1,3) = .813$, $p = .434$, $\eta2 = .213$. Finally, there was no three-way interaction, $F(1,3) = 1.161$, $p = .360$, $\eta2 = .279$.

Qualitative Analysis

In addition to the statistical analysis, qualitative analysis was completed to see where trends may be occurring that warrant further investigation in future studies with a larger number

of participants. The following figures show the average results by session and by recording time (baseline, feedback) by gender with standard error bars to indicate the variability within genders. Please note that standard error bars are not present for session 8 for the men. This is due to only one male instructor being recorded for the eight session, so no error bars were generated.

Pitch Strength

The summary results can be seen in Figures 38 and 39. Figures 38 and 39 show separate linear regressions for pitch strength and first moment specific loudness, respectively. Unlike Part 2.1 where opposite trends were found for men and women, the same basic trend emerges when looking at Figure 39. For both genders, there is a decrease in change pre- to post-VLT for both men and women. This trend, although not statistically significant, follows the hypothesized direction.

Figure 38: Average change scores (pre – post) for pitch strength by session in Part 2.2. Positive values indicate greater pitch strength before the vocal loading task.

In addition, this decrease indicates less of a decrease in pitch strength over the course of the

VLT, leading to better voice quality (less fatigue) post-VLT. This trend is also seen in the

session-to-session comparison in Figure 29. For men, the decrease appears more gradual. For

women, there is a steep decline between feedback sessions 2 and 3.

Figure 39: Average change scores (pre – post) for pitch strength by recording type in Part 2.2.
Positive values indicate greater pitch strength before the vocal loading task.

First Moment Specific Loudness

The summary results can be seen in Figures 40 and 41. Figures 40 and 41 show separate

linear regressions for pitch strength and first moment specific loudness, respectively. Unlike

pitch strength in Part 2.2 where both genders showed the same trend, opposite trends were found

for men and women. Women followed the hypothesized trend of a decrease in change in first

moment specific loudness (-0.17 to 0.03), but men showed the opposite trend (-0.26 to -0.37).

While a negative value indicates an increase in first moment specific loudness (strain) post-VLT,

the switch to an average positive value for women in the feedback recordings indicates a

reduction in strain post-VLT.

Figure 40: Average change scores (pre – post) for first moment specific loudness by session in Part 2.2. Positive values indicate greater first moment specific loudness before the vocal loading task.

Figure 41: Average change scores (pre – post) for first moment specific loudness by recording type in Part 2.2. Positive values indicate greater first moment specific loudness before the vocal loading task.

The pitch strength and first moment specific loudness findings for women in Part 2.2 followed the hypothesized direction, suggesting that both of these may be possible outcome measures for female occupational voice users. However, with a small group of three individuals,

this trend must be interpreted with caution. These findings do suggest that further exploration of these measures as outcome measures may be appropriate for female occupational voice users.

Summary

Correlations between accelerometer and audio recordings for the auditory measures (pitch, pitch strength, first moment specific loudness) were conducted because these measures have previously been validated only for use with audio recordings (Kopf et al., 2013; Shrivastav & Camacho, 2010; Shrivastav et al., 2012). A high correlation (0.901) indicated that pitch estimates from Auditory-SWIPE' are appropriate to use with both accelerometer and audio recordings. This measure also correlated with fundamental frequency estimates at 0.961. However, these recordings were of individuals with normal, Type 1 voices, and the predicted correlations for Type 2 and Type 3 voices would be reduced (Shrivastav et al., 2014). The other two measures are not appropriate to analyze from accelerometer recordings due to low correlations.

For Hypothesis 1, Part 2.1 found a statistically significant difference in pitch between men and women, with men having a lower pitch. These findings are consistent with the previous literature. Both Part 2.1 and Part 2.2 found a statistically significant difference of measure, which was also anticipated due to the measures addressing different constructs and having different scales. While no other statistically significant differences were found, similar trends were seen across the two phases of the study for phonation time and pitch by gender. Men in both phases showed a decrease in phonation time and pitch with feedback, both in the hypothesized direction. On the other hand, women in both phases showed increases in phonation time and pitch with feedback. This finding suggests that there may be a gender-specific response to the feedback that

should be further examined with a larger pool of participants. In addition, feedback may need to be tailored differently for women to encourage the same risk-reducing changes found in men. Finally, participants in Part 2.1 showed an overall increase in dB level in the feedback recordings, whereas participants in Part 2.2 showed an overall decrease (the hypothesized direction). This discrepancy may be due to the difference in VLTs, with Part 2.1 participants instructed to speak at a louder than normal volume.

For Hypothesis 2, there was a significant, although small, finding that pitch strength is a predictor of perceive vocal fatigue. While most participants reported some increase in fatigue, one participant (P2101) reported no fatigue at any point in the study, and participant-rated fatigue often increased by only 1 or 2 points on a 10-point scale after the VLT. However, looking at the data, one can see the high variability of associations between ratings and pitch strength. This is not exclusive to ratings of vocal fatigue, and has been observed, for example, in clinician ratings of voice quality (Kreiman et al., 1993). Possible ways to reduce this variability, and therefore get a more robust rating to compare with objective measures, would be to standardize and average multiple ratings (Shrivastav et al., 2005), which would need to be done systematically in a future study.

Finally, for Hypothesis 3, there was no statistically significant change in pitch strength or first moment specific loudness over time with the VLTs. However, when looking at the trends, women in Part 2.2 (current occupational voice users) showed the expected trends for both pitch strength and first moment specific loudness, with men also showing the expected trend for pitch strength only.

In Part 2.1, men did show a reduction in pitch strength change scores in the feedback recordings, but they showed an increase in pitch strength (increase in quality post-VLT) in the

baseline recordings and the opposite in the feedback recordings. While women had an increase in the amount of change in the feedback recordings, this change demonstrated an increase in quality post-VLT with feedback whereas these individuals had a decrease in quality post-VLT in the baseline recordings. The exact opposite findings were discovered for first moment specific loudness. Based on Part 2.1 findings, the direction and magnitude of change did not move together in the hypothesized direction.

Conclusions

Overall, the statistical findings partially support the current hypotheses. Correlational analyses support the use of pitch from accelerometer recordings, but do not support calculations of pitch strength and first moment specific loudness. In addition, pitch strength was found to be a predictor of vocal fatigue in Part 2.1.

Qualitative analysis of the data provided a richer context for exploring the trends. These trends suggested that men decreased phonation time and pitch in response to feedback, both of which reduce the risk of developing voice disorders. Findings from Part 2.2 participants indicated that, regardless of gender, individuals reduced vocal intensity with feedback, again reducing voice disorder risk. In addition, Part 2.2 participants demonstrated a decrease in pitch strength change scores pre- to post-VLT, indicating less change (possibly less fatigue) as a result of feedback. In addition, all of these average change scores were positive (pitch strength decreasing over the VLT), and this reduction in change indicates less of a decline in voice quality with feedback. Further analysis with a larger set of participants should be completed to determine if any of these findings may be statistically significant.

CHAPTER 6: Study Discussion and Conclusions

Summary of Findings

The goal of the current study was to determine whether Preventative Voice Monitoring (PVM), using a dosimeter with task-based feedback, impacts vocal behavior in occupational voice users and future occupational voice users. Three study aims were defined to design the feedback (Aim 1) and to assess its ability to influence both behavior change measures (Aim 2) and voice production (Aim 3). To address the aims, the study was divided into two parts: the creation of the feedback (Part 1) and testing of the feedback (Part 2). Summarized findings are included below.

Aim 1: At the end of Part 1, a layered display structure for four measures (loudness, pauses, quality, strain) was developed with three displays with varying views of the data. The first display was an "at a glance" version, the second display allowed comparison across multiple days, and the third display allowed a more detailed view of an individual day's data. Participants in Part 2 suggested further improvements to future iterations of the feedback, including better defined measures, suggested behavior changes based on feedback results, and the ability to provide further context for later reflection (e.g., links to calendar and ability to add personal notes on the voice).

Aim 2: A statistically significant increase in questionnaire scores was found after PVM (both at baseline and with feedback). These increases were positive for RTC and S-E, with an increase in these scores indicating positive behavior change. For the VFI, only an increase in Factor 3 demonstrates an improvement. However, increases in Factors 1 and 2 may indicate increased awareness of fatigue, which may later lead to increased desire for behavior change.

Aim 3: There were some statistically significant findings related to the three hypotheses. For Hypothesis 1 (phonation time, pitch, and/or vocal intensity will decrease with PVM), there was a main effect of measure in Parts 2.1 and 2.2, indicating that the measures were quantifying three distinct attributes. There was an interaction of gender and measure in Part 2.1, with post hoc testing revealing that men had a higher average pitch than women (consistent with prior literature). For Hypothesis 2 (pitch strength and first moment specific loudness can predict vocal fatigue), stepwise linear regression found a small but statistically significant result that pitch strength was able to account for 2.8% of the variance in vocal fatigue self-ratings. Finally, for Hypothesis 3 (the change in pitch strength and first moment specific loudness from before to after the VLT will decrease with feedback), there were no statistically significant results.

Despite few statistically significant results for Aim 3, there were a number of trends identified that warrant future exploration. For Hypothesis 1, average phonation time increased for women in both parts of the study and men in Part 2.1 with feedback, but decreased for men in Part 2.2 (hypothesized direction). For average vocal intensity, both genders in Part 2.1 showed an average increase with feedback, while both genders in Part 2.2 showed an average decrease (hypothesized direction). All women showed an average increase in pitch with feedback, and men showed a decrease (hypothesized direction).

For Hypothesis 2, participants in Part 2.1 showed an average decrease in both pitch strength (hypothesized direction) and first moment specific loudness with increasing vocal fatigue. On the other hand, participants in Part 2.2 showed an increase in pitch strength and virtually no change in first moment specific loudness with increasing vocal fatigue.

For Hypothesis 3, Part 2.1 pitch strength trends demonstrated an interaction of gender and time. While only men decreased the change in pitch strength with feedback (hypothesized

direction), men changed from having higher pitch strength post-VLT to pre-VLT, and women had the opposite trend. In Part 2.2, both genders saw an average decrease in change in pitch strength with feedback (hypothesized direction), and all change scores were positive (indicating that pre-VLT scores were higher than post-VLT scores; hypothesized direction).

Women in both phases of Part 2 showed reduced pre-VLT to post-VLT change in first moment specific loudness with feedback (hypothesized direction), while men showed increased change. However, opposite trends for the direction of change were seen between genders in both phases. In Part 2.1, women saw a decrease in first moment specific loudness at baseline and an increase with feedback (hypothesized direction). Men saw the opposite trend from the women. In Part 2.2, women saw an increase in first moment specific loudness at baseline, but a decrease with feedback. Again, men showed the opposite trend from the women.

Study Limitations

While some statistically significant results were found, it is important to interpret the significance with caution due to the small number of participants in each group, especially the group in Part 2.2. However, despite these small numbers, some large effect sizes were found, which strengthens the likelihood of these findings remaining consistent in future studies. Future work in this area should include larger groups in order to increase the confidence in the statistical results.

In addition to statistical limitations, other limitations include the VLTs. In Part 2.1, the task was artificial (reading aloud to an imaginary classroom), and a number of behavioral changes reported were attributed to the task itself. Another important point is that the participants in Part 2.1 were predominantly students who had limited experience as occupational voice users,

and this was reflected in many of their conjectures of how feedback would be different for the target population.

In Part 2.2, the VLT was an actual job-related task (course instruction) performed by occupational voice users. However, this was also limited by factors that are inherent to doing in situ research. First, there was great variability between the courses taught by the instructors in length, content, and class size. This variability may have contributed to some of the non-statistically significant findings, in addition to the small sample size. Furthermore, these individuals were monitored for only a short portion of their work day. Monitoring these individuals in multiple settings and multiple contexts is important for future assessment of the feedback system, to assess its flexibility and other user needs that may arise in different contexts.

One participant, P2109, reporting shifting his focus from observing internal cues about his voice (baseline recordings) to ignoring internal cues in favor of external cues (the feedback). While only one participant reported this change, it is interesting to consider in light of further PVM development. In voice therapy, clinicians not only focus on teaching patients to produce a better voice, but also teaching patients to focus on internal cueing of better voice production (Ramig & Verdolini, 1998). The idea behind PVM is to increase awareness of the voice, and provide a way for individuals to monitor the voice over time to try to optimize production before the need for treatment arises. Therefore, future studies need to explore the best combination of objective and subjective information for PVM to support self-reflection on voice production.

Another limitation of the current study is the organization of the study. Based on participant interview data, participants spent time figuring out how measures were related, what measures meant, and only some attempted to change their voice based on the feedback. In future

studies, a trial period with the feedback to see how it works may allow more future participants to focus on voice behavior change.

Feedback Issues

Sometimes recording or analysis issues led to incomplete feedback being presented to participants for a particular session. However, for most participants in Part 2.1, only one or two sessions had incomplete feedback (only one or two of the three measures presented). For P2111, this unfortunately happened on multiple occasions due to an undetermined issue with the recorder where the initial and final tasks were usually usable, but the VLT information from the accelerometer was unreliable. Because of the issues with many of her recordings, P2111 received limited feedback. For P2106 and 2109, part of one day's recording was cut off at the end, but this did not happen for other participants recorded using the same equipment on the same day before and between these participants (the issue was not discovered until all the day's recordings were completed).

In Part 2.2, two participants had recurrent recording issues. P2202 nearly always had a background noise of approximately 78 Hz present through the majority of the accelerometer recordings, which is believed to be attributed to interference from other electronic devices in the classroom itself. The initial and final tasks were usually clean (performed in the laboratory), but the classroom portion of the recording usually contained the noise. Therefore, this participant received feedback on quality from the initial and final Rainbow Passage readings rather than from the classroom recording (for post hoc analyses, we used an 80 Hz high pass filter to get rid of the extraneous noise). For P2204, the first two classroom recordings were lost due to a loose

battery issue that led to intermittent loss of recordings. One later classroom recording was lost due to an unknown recorder error.

Another somewhat consistent issue with recordings in Part 2.2 was the presence of long periods of silence during the lecture (due to student presentations, videos, guest speakers, etc.). If one of these periods occurred in the first or last 15 minutes of the course, pitch strength values were not reported in feedback during these silences. Occasionally, there would also be moments during the lecture where the pitch strength would drop by 0.2 or more for about a minute and then recover to the initial level for the remainder of the segment. Sometimes this was associated with increased silence duration, or the presence of temporary noise, but with no appreciable difference heard in the speaker's voice quality perceptually. These data points were also omitted in the feedback.

Another potential feedback issue was the measures used in feedback. While these measures were identified as salient to occupational voice users in Part 1, participants in Part 2 reported some frustration with little to no change in these measures over the course of the study. This limited range may be due to a ceiling effect because these measures, especially the voice quality measures, are used in differentiating between normal and dysphonic voices. Participants in this study fell within the normal range of voices, so there was potentially little room for change on these measures. Therefore, future research should investigate the measures used in feedback, and look at measures that have been shown to differentiate between groups of non-dysphonic voices. Some possible measures to examine in future studies include measures from the long-term average spectrum, standard deviation of sounded (voiced) durations, and smoothed cepstral peak prominence (Warhurst et al., 2016; Warhurst, McCabe, Yiu, Heard, & Madill,

2013). These measures have been identified in studies to differentiate between commercial radio voices and both public radio and control voices in Australia.

Finally, biofeedback such as that provided in physical activity tracking instructs users to change behavior in a singular direction (e.g., increasing the number of steps taken). For PVM, the goal of the feedback is to help occupational voice users to find a balance in voice use. For example, users should decrease phonation time to give the voice more chance to rest, but still maintain a level of phonation time for conveying their message. Therefore, future research should evaluate whether there are better means of conveying this type of information to users.

Future Implications

During the study, participants provided a vast amount of helpful input. While there were few changes in objective voice measures across participants, input provided in the interviews will inform future design. Suggestions for future exploration include: better understanding of the needs of the array of potential users, the inclusion of recommendations based on feedback, including notes options to allow better comparison, and providing a social component to share data or potentially just advice. While students had mixed opinions on whether they would use the future app, current occupational voice users expressed interest in the app after appropriate modifications are made.

Conclusions

Overall, the results of this study support the further iteration and eventual design of a mobile application for occupational voice users to prevent voice disorders. The findings from the study indicate that users want an interactive, flexible system with well-defined measures and a

layered display structure that provides both objective feedback and suggestions to improve voice use. This insight from P2114 sums up the purpose of the study nicely: "I really liked the study and I feel that the more people know about their voices, the less voice problems people will have." This sentiment suggests that PVM could fill an important void for occupational voice users who may want to change their voices but do not feel that they have the knowledge or skills to make the changes on their own. Preventative Voice Monitoring could be used to empower occupational voice users and help them avoid future voice disorders.

APPENDICES

Appendix A: Intake Form (Modified VBALAB form)

Gender: M   F   Age: _____

a. Compared to a normal day, today I feel:

   a) Much less stress

   b) Less stress

   c) The same stress

   d) More stress

   e) Much more stress

b. Compared to a normal day, today I feel:

   a) Much less fatigue

   b) Less fatigue

   c) The same fatigue

   d) More fatigue

   e) Much more fatigue

c. I would describe my PRIMARY

workplace as:

   a) Very Quiet

   b) Quiet

   c) Neutral

   d) Noisy

   e) Very Noisy

e. Do I commonly experience symptoms of reflux (or heartburn)?   yes   no
    Am I experiencing reflux symptoms today?   yes   no

f. Do I commonly experience symptoms of seasonal allergies?   yes   no
    Am I experiencing allergy symptoms today?   yes   no

g. In the past year, I have smoked: never   occasionally   daily

h. On average, I consume ___ caffeinated beverages per day.
   0   1   2   3   4+
i. Do you use voice amplification (e.g., microphone) in your job?
   ___ Yes   ___ No

It is important that the ethnic and racial makeup of our research participant pool reflects that of the local community. Please indicate which of the following ethnic and racial categories you identify:

Ethnic Category	
Hispanic or Latino	
Not Hispanic or Latino	
Prefer not to identify	

Racial Categories	
American Indian/Alaska Native	
Asian	
Native Hawaiian or other Pacific Islander	
Black or African American	
White	
Prefer not to identify	

Table 13: Ethnic category.   Table 14: Racial category.

1. Are you a native speaker of American English?

___ Yes   ___ No

2. Do you have, or have you ever had a voice disorder that required treatment (therapy, surgery)?

___ Yes   ___ No

3. Do you have any hearing loss?

___ Yes   ___ No

4. Do you have, or have you ever had a speech disorder that required treatment (therapy, surgery)?

___ Yes   ___ No

5. Do you use voice amplification (e.g., microphone) in your job?

___ Yes   ___ No

Appendix B: Initial Semi-Structured Interview (Parts 1 & 2)

1. Tell me about your job (if occupational voice user).
   a. Do you talk/sing a lot as part of your job?
   b. How noisy is your work environment?
   c. Tell me about your voice use during a typical work day.
   d. About how many hours do you talk per day in your job?
2. Tell me about your future career goals (if participant is a student).
3. Have you ever experienced any problems with your voice?
   a. Have you ever been sick and lost your voice?
   b. Do you ever feel like your voice gets tired at the end of the day? When?
   c. The last time that your voice became tired, what did you do? Was it helpful?
   d. Have you ever sought help for these problems?
      i. Who did you turn to?
      ii. What did you do?
   e. Do you have any current concerns about your voice?
4. Have you ever heard that certain professionals, such as teachers, are at risk for voice disorders?
   a. Where did you hear this?
   b. What did you think?
   c. Do you know of anyone in your profession or another profession who has had voice disorders? Who? What action did they take?

Appendix C: Part 1 Semi-Structured Interview Questions

After all measures of interest are explained, the researcher will ask the following:
1. How important do you think this measure is to understanding your voice on a scale from 1-10, where 1 = not at all and 10 = extremely? Why?
2. How useful do you think this measure could be as daily feedback on a scale from 1-10, where 1 = not at all and 10 = extremely? Why?
3. How confident are you that you could change this measure based on daily feedback on a scale from 1-10, where 1 = not at all and 10 = extremely? Why?

For each measure, the following question will be asked:
1. If you were to receive feedback on this voice measure, what should it look like? You can tell me in words or I have a piece of paper you can draw on.

The participant will be asked to look at all three visual displays for a given measure first sequentially and then simultaneously and answer the following:
Sequential Viewing (Mazza, 2006)
4. What kind of information can you gather from this feedback display?
5. Can you identify general patterns or tendencies in the feedback display?
6. Is the feedback display useful? Why?
7. Do you think the feedback should be presented differently? How?
Simultaneous Viewing
1. Please tell me which of these displays would be most helpful for you. Why?
2. Please tell me which of these displays would be least helpful for you. Why?
3. Are there any parts from these visual displays that you would combine to create a better visual display? Which parts? Why?
4. (Taking out the earlier description/drawing of feedback display) How do these compare to your initial thoughts on the feedback display?
   a. Which do you like better?
   b. Do you think they could be combined?
   c. Now that you've seen these prototypes, could you tell me what your ideal feedback display would be?
   d. Is there anything else that might be useful that isn't captured in this feedback display? (Mazza, 2006)

Appendix D: Initial Feedback Displays

Distance
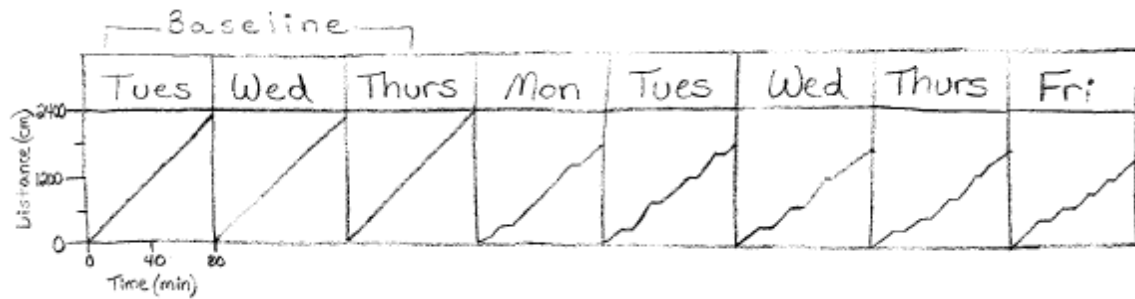
Figure 42: Initial displays- Distance small multiple line graph.
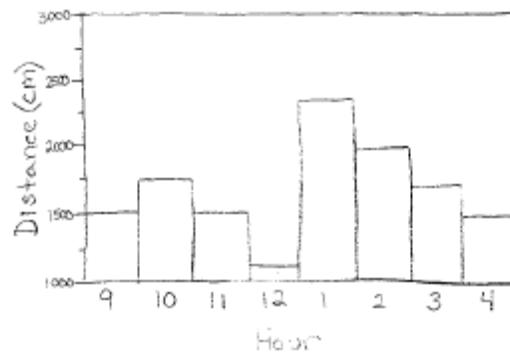
Figure 43: Initial displays- Distance bar graph.
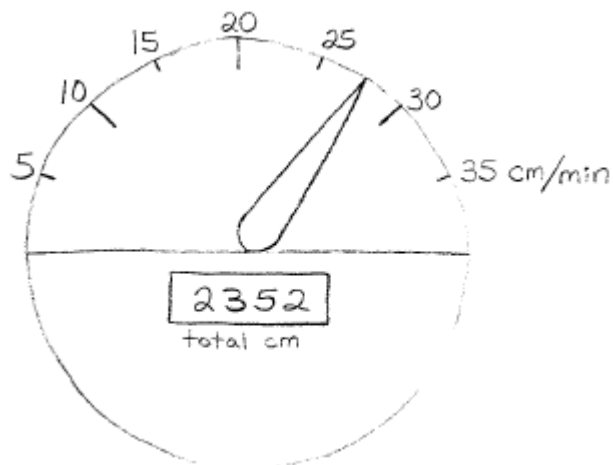
Figure 44: Initial displays- Distance speedometer

Loudness
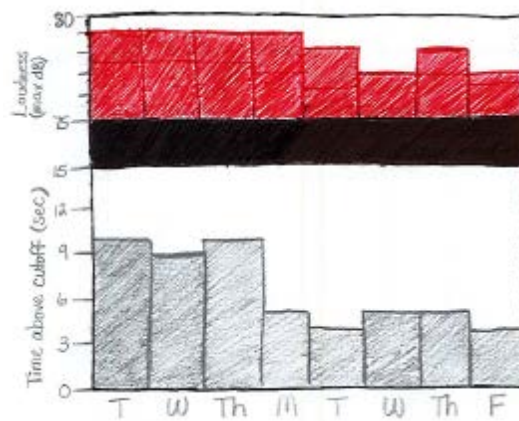
Figure 45: Initial displays- Loudness bar graph.
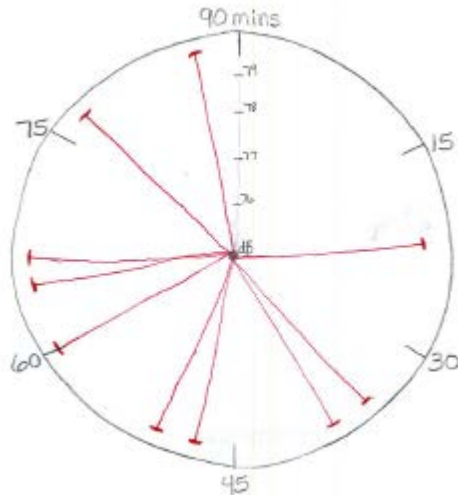
Figure 46: Initial displays- Loudness clock.
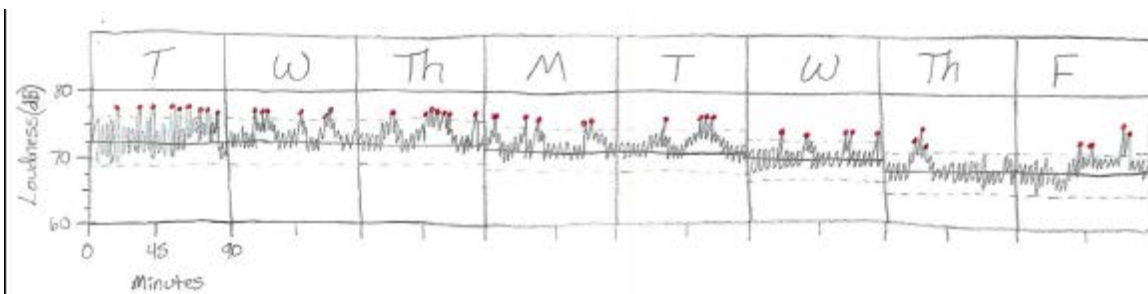
Figure 47: Initial displays- Loudness small multiple sparklines.
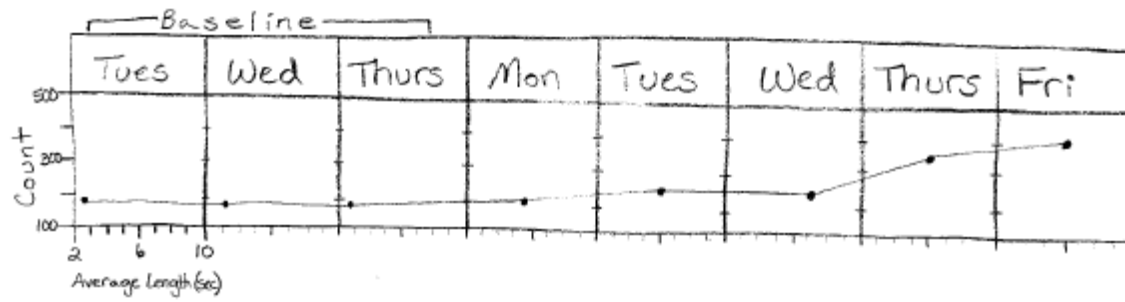
Pauses

Figure 48: Initial displays- Pauses small multiple line graph.
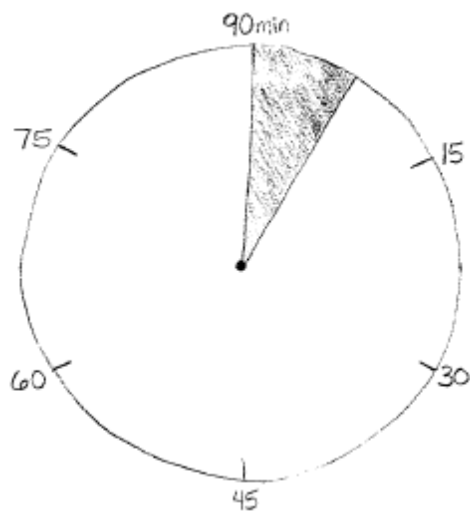
Figure 49: Initial displays- Pauses clock.
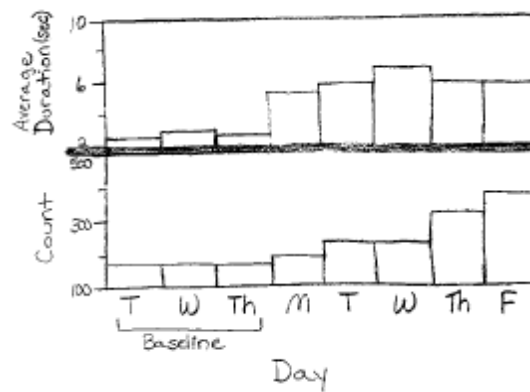
Figure 50: Initial displays- Pauses bar graph.
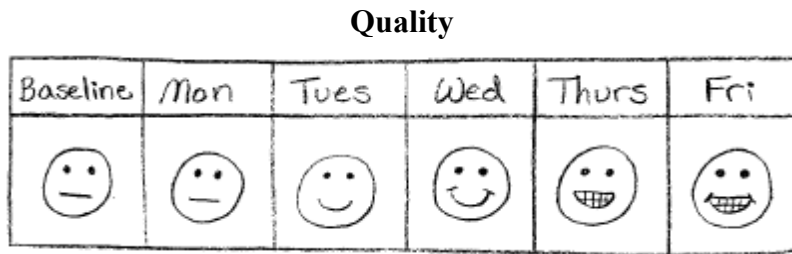
Quality

Figure 51: Initial displays- Quality small multiple smileys.
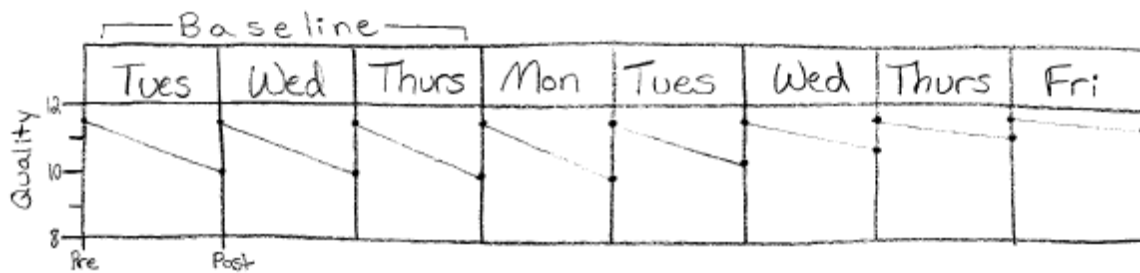
Figure 52: Initial displays- Quality small multiple line graphs.
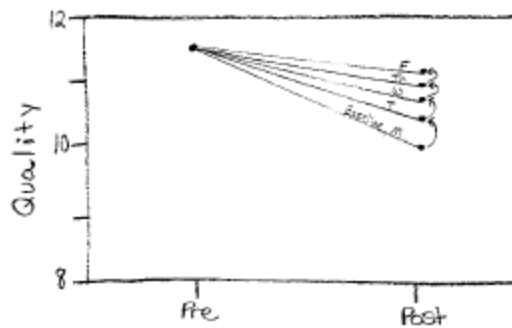
Figure 53: Initial displays- Quality multi-day line graph.
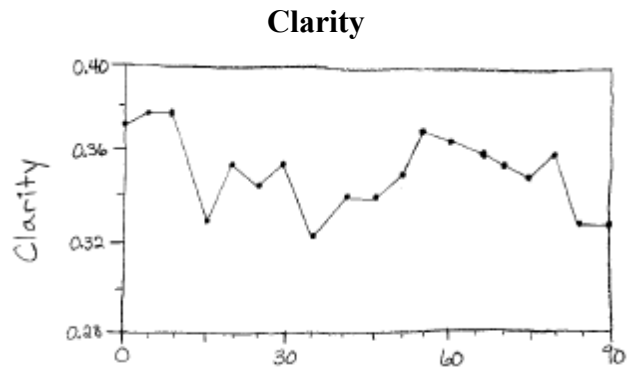
Clarity

Figure 54: Initial displays- Clarity single-day line graph.
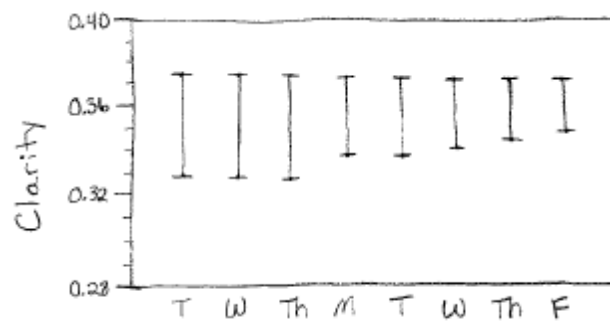
Figure 55: Initial displays- Clarity vertical line graph.
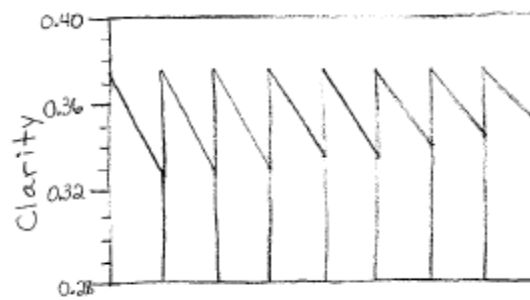
Figure 56: Initial displays- Clarity bar graph.

Strain
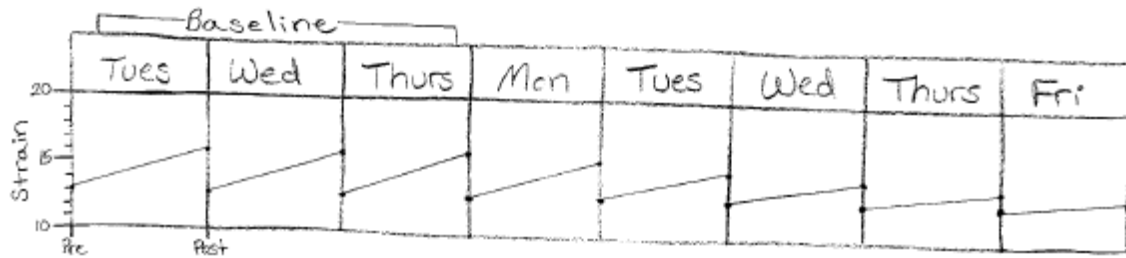
Figure 57: Initial displays- Strain small multiple line graph.
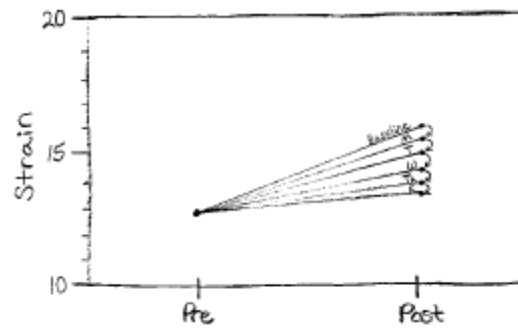
Figure 58: Initial displays- Strain multi-day line graph.
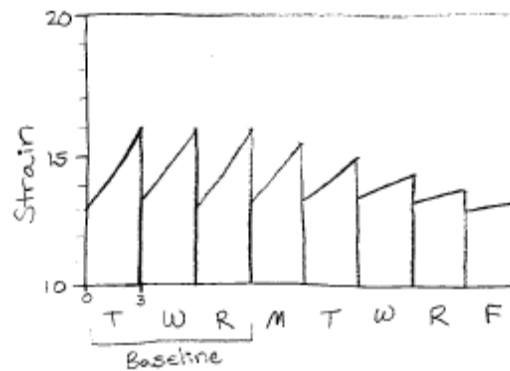
Figure 59: Initial displays- Strain bar graph.

Multi-Measure

Figure 60: Initial displays- Multi-measure matrix.

Appendix E: Iteration 1 Feedback Displays
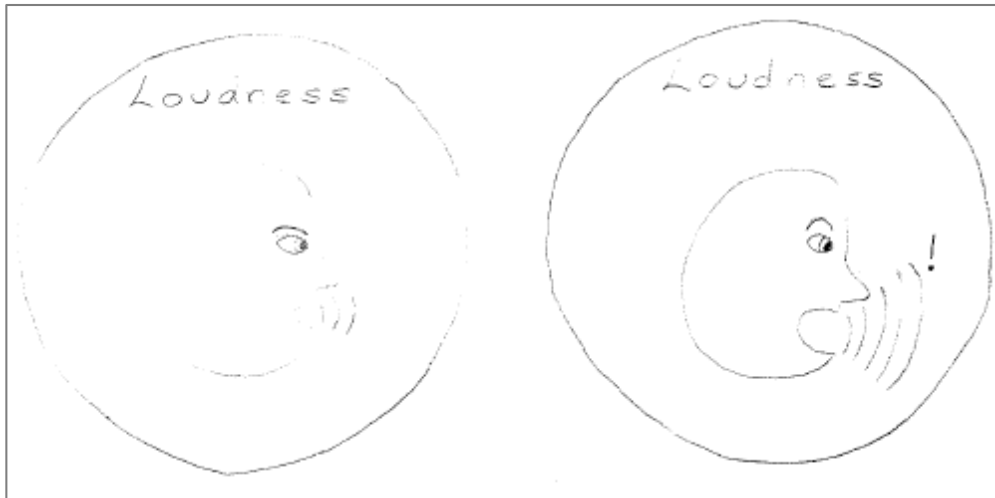
Icons

Figure 61: Iteration 1 displays- Dynamic loudness icons.
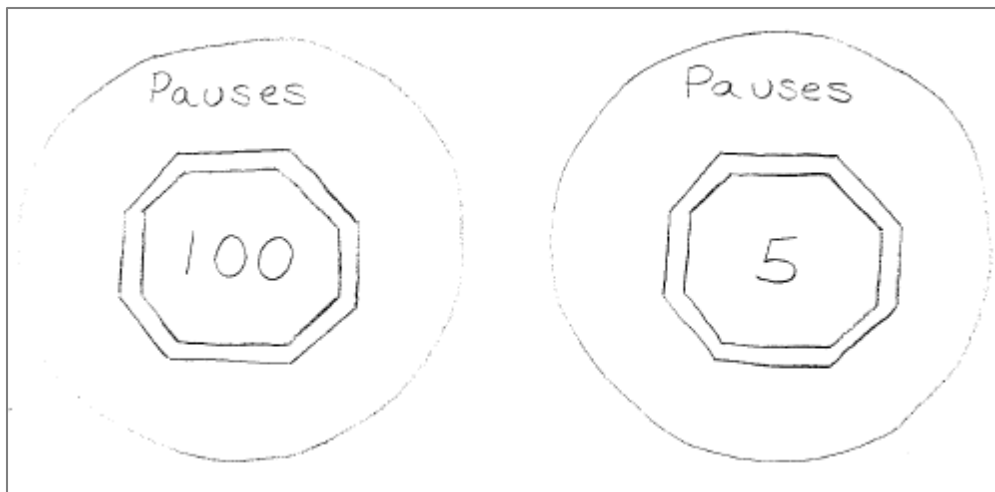
Figure 62: Iteration 1 displays - Dynamic pause icons.

Icons
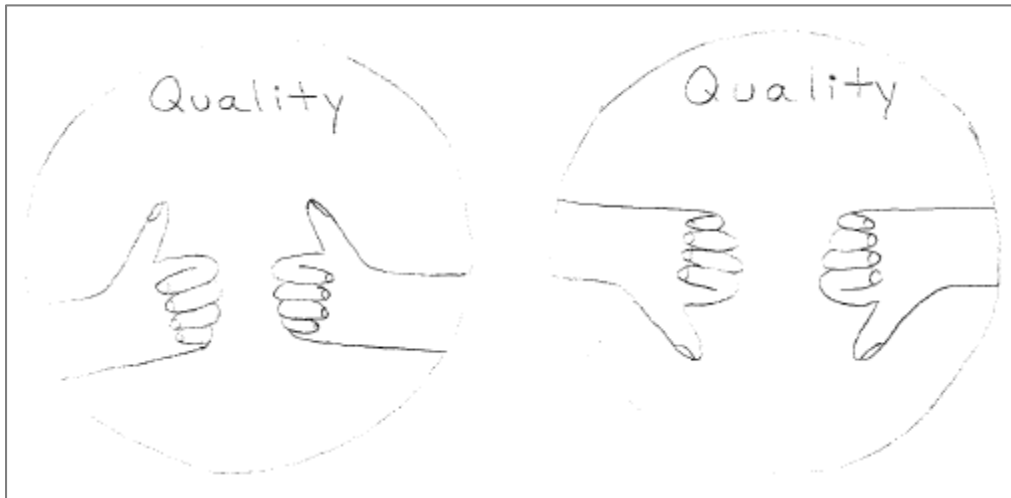
Figure 63: Iteration 1 displays - Dynamic quality icons.

Figure 64: Iteration 1 displays - Dynamic strain icons.
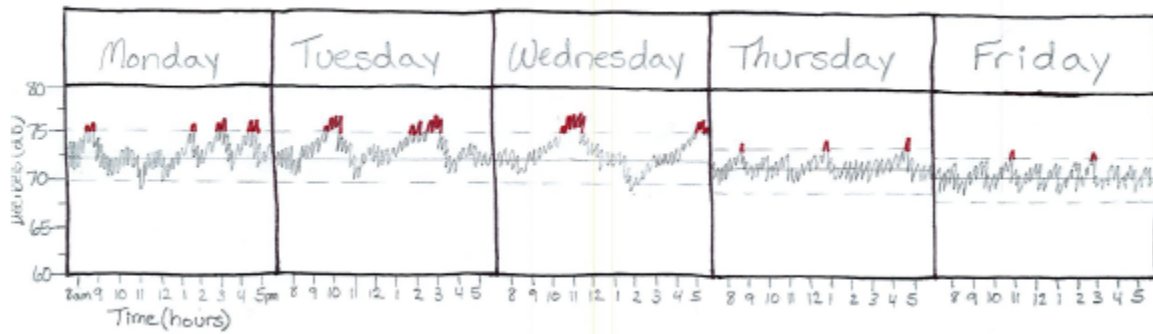
Loudness
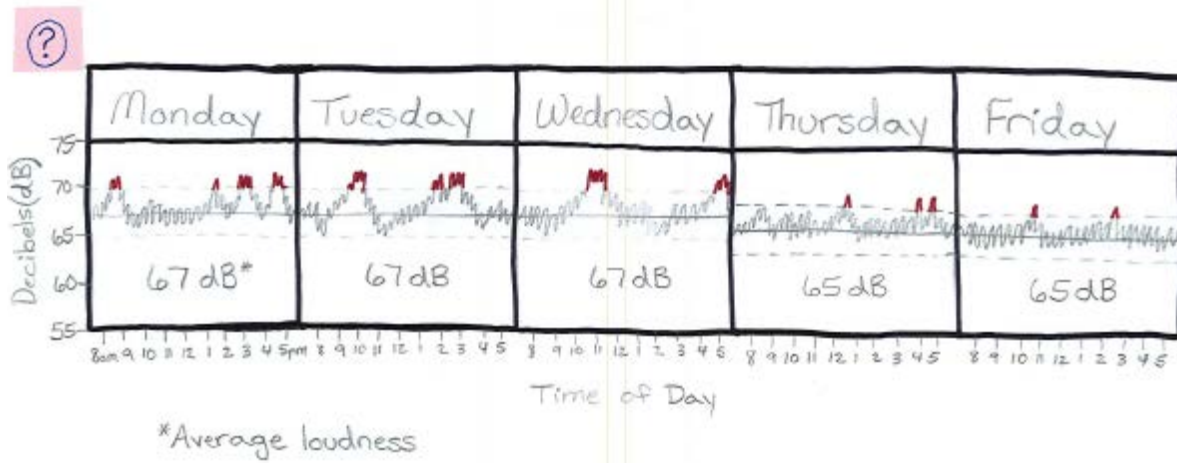

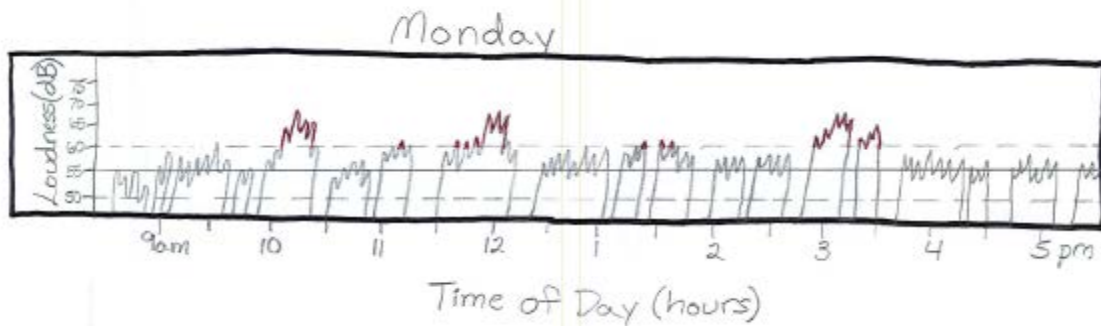Figure 65: Iteration 1 displays – Loudness small multiple sparklines.
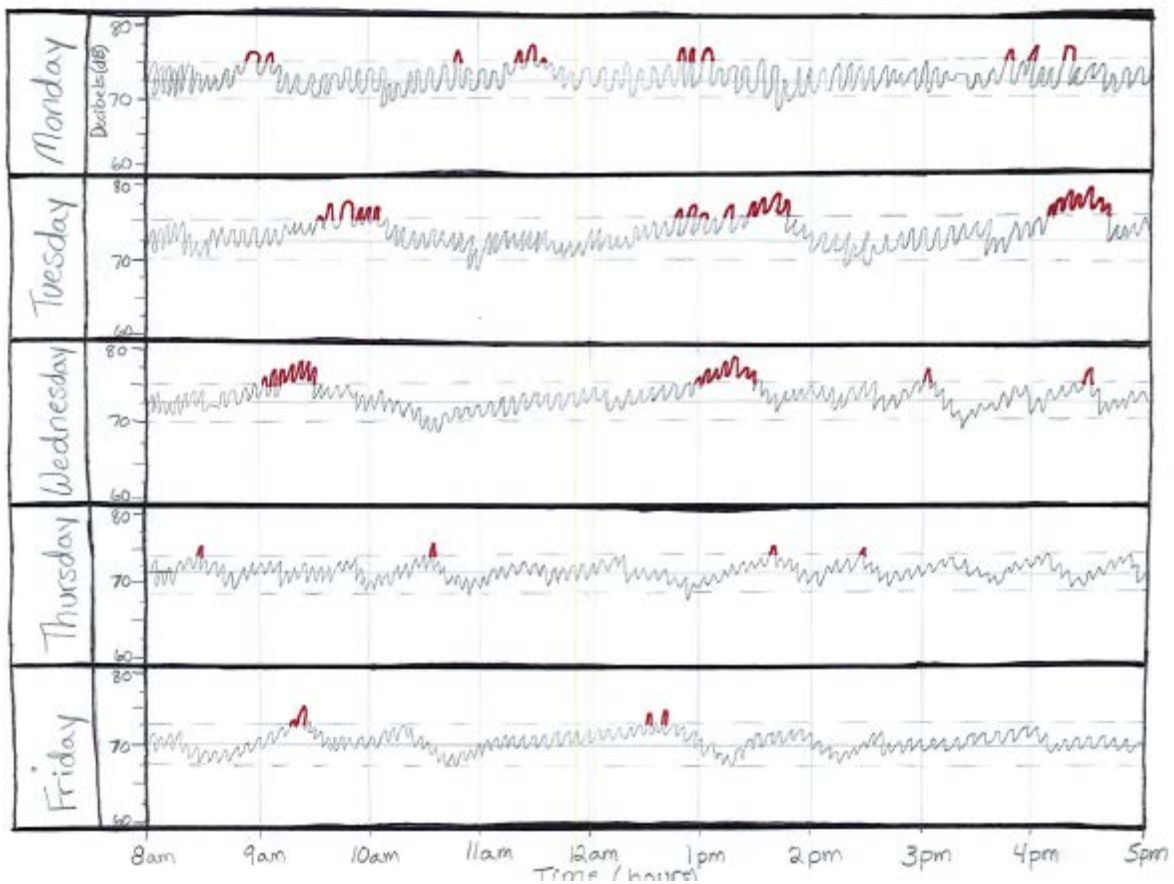
Figure 66: Iteration 1 displays – Loudness multiple sparklines.
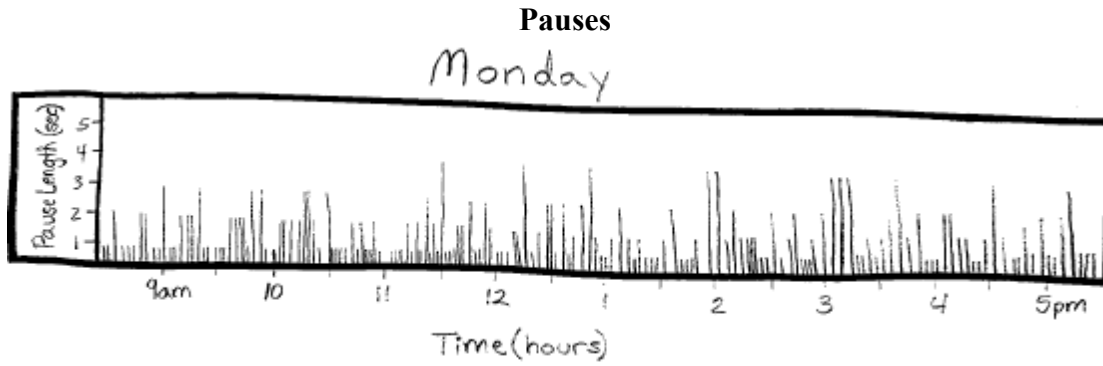
Pauses

Figure 67: Iteration 1 displays – Individual pause time and length.
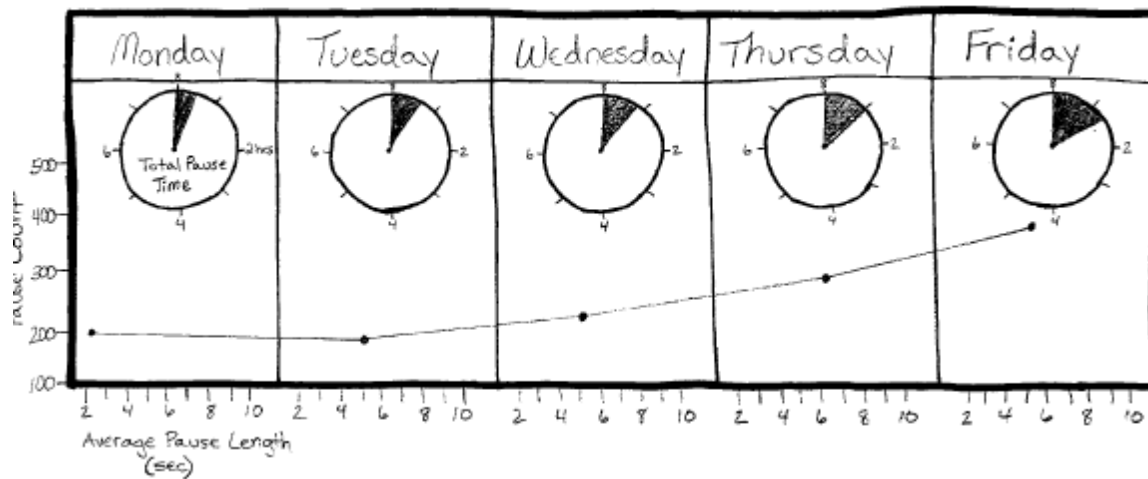
Figure 68: Iteration 1 displays – Pause clocks and line graph.
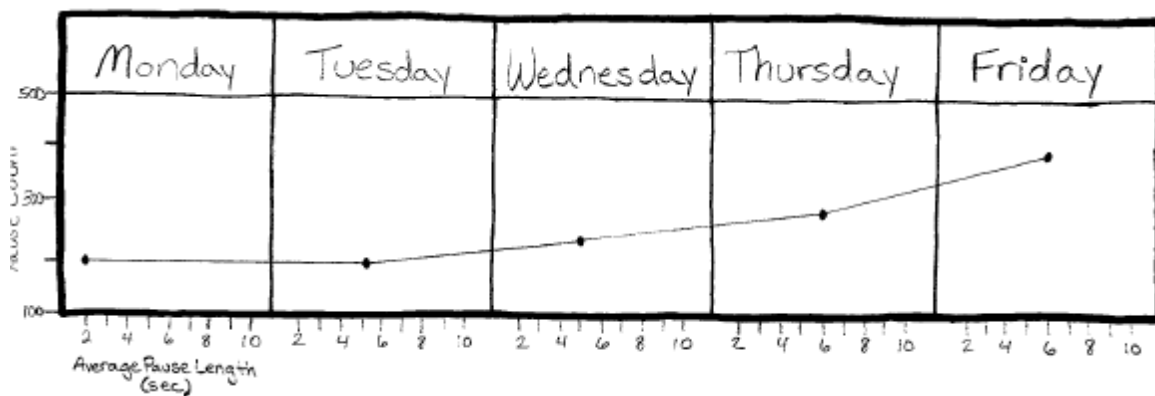
Figure 69: Iteration 1 displays – Pause line graph.

Quality

| Baseline	Mon	Tues	Wed	Thurs	Fri
😐	😐	🙂	😊	😁	😁

Figure 70: Iteration 1 displays – Quality small multiple smileys.
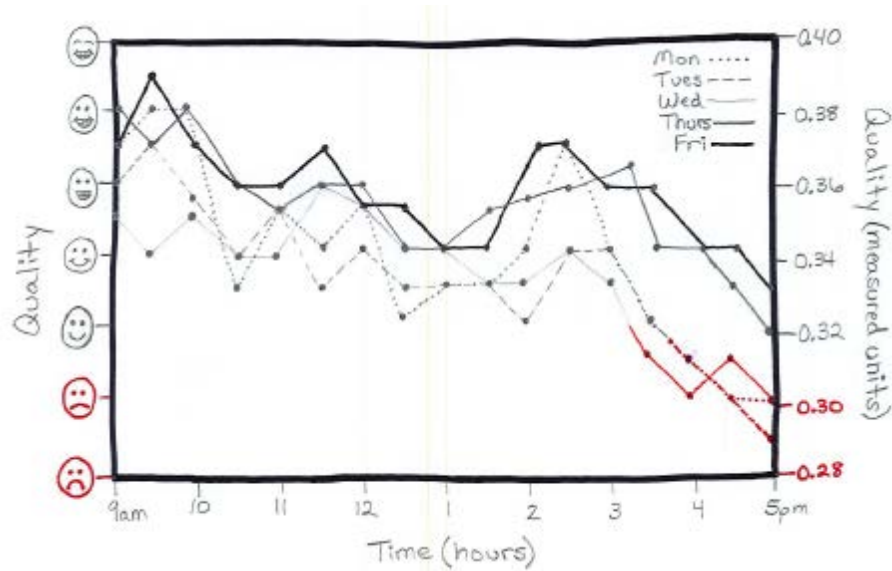
Figure 71: Iteration 1 displays – Quality multi-line graph.
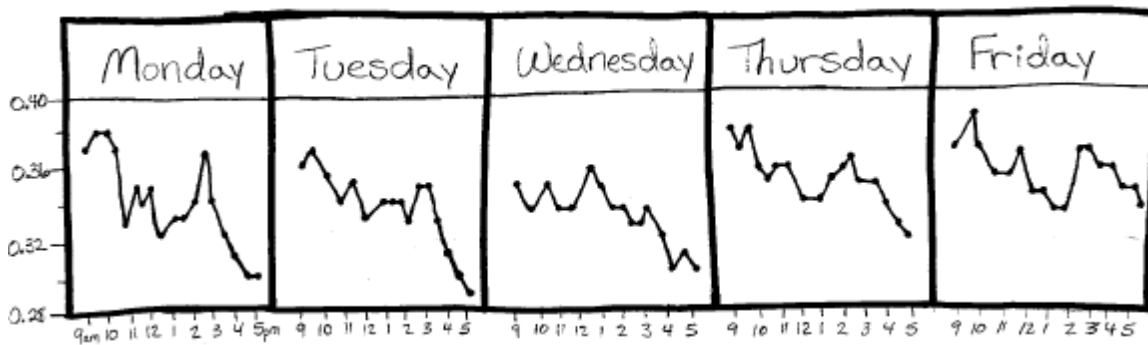
Figure 72: Iteration 1 displays – Quality small multiple line graph.
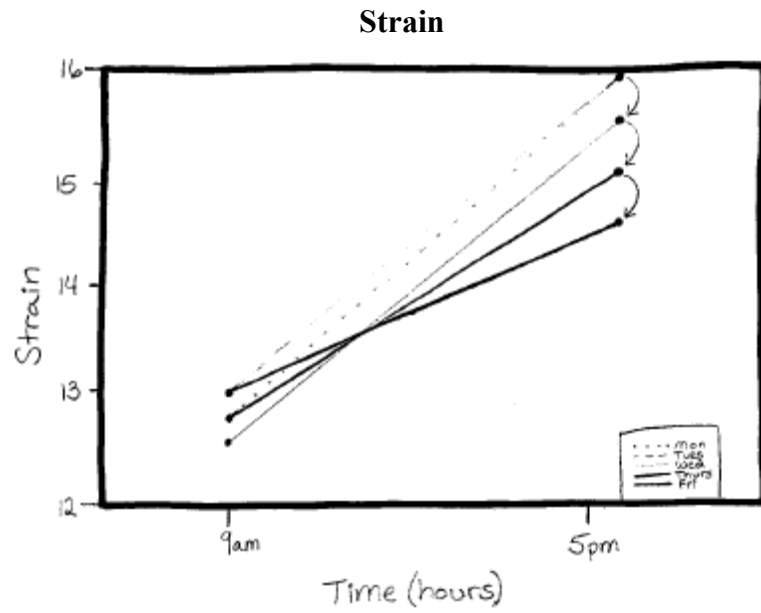
Strain

Figure 73: Iteration 1 displays – Strain multi-shade line graph.
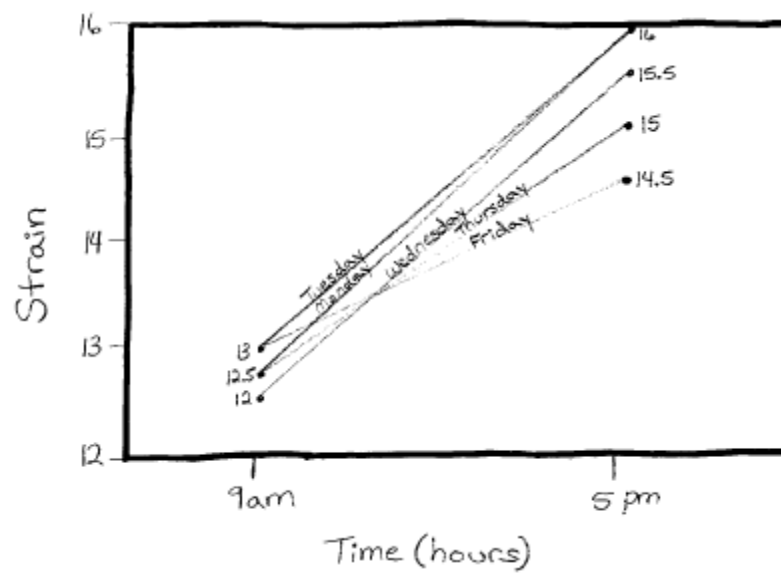
Figure 74: Iteration 1 displays – Strain labelled line graph.

Multi-Measure

Figure 75: Iteration 1 displays - Multi-measure matrix.

Appendix F: Iteration 2 Feedback Displays

Icons

Figure 76: Iteration 2 displays- Dynamic loudness icons.
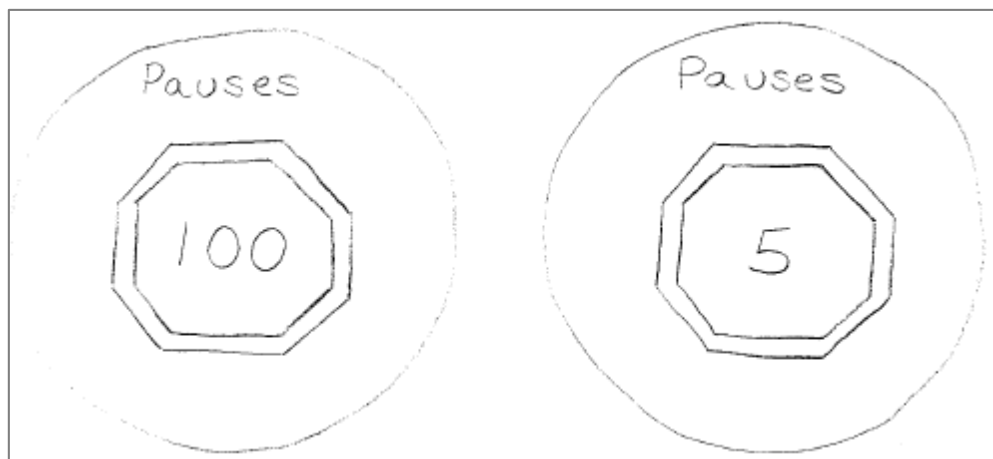
Figure 77: Iteration 2 displays - Dynamic pause icons.
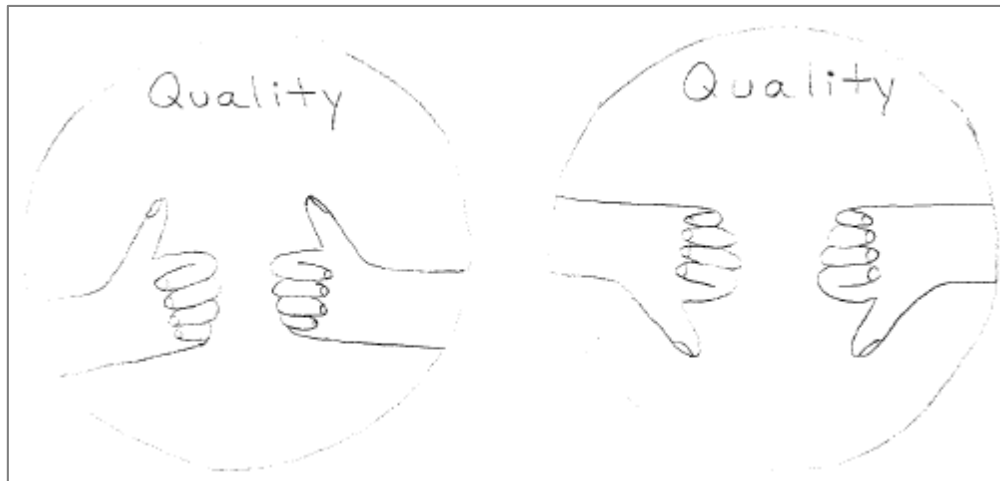
Figure 78: Iteration 2 displays - Dynamic quality icons.

Figure 79: Iteration 2 displays - Dynamic strain icons.
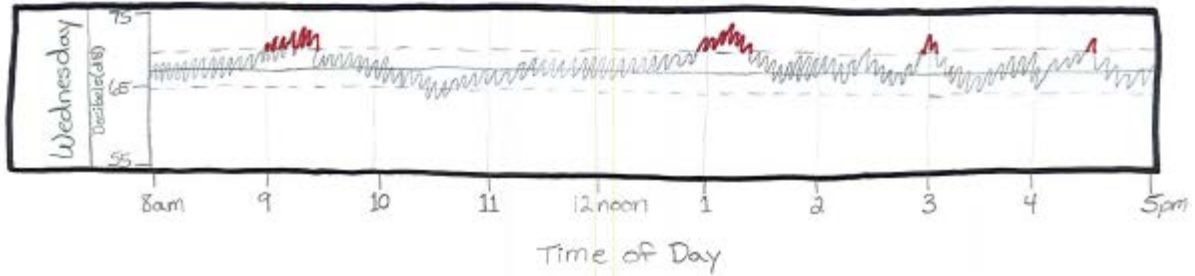
Loudness

Figure 80: Iteration 2 displays – Loudness single-day sparkline.
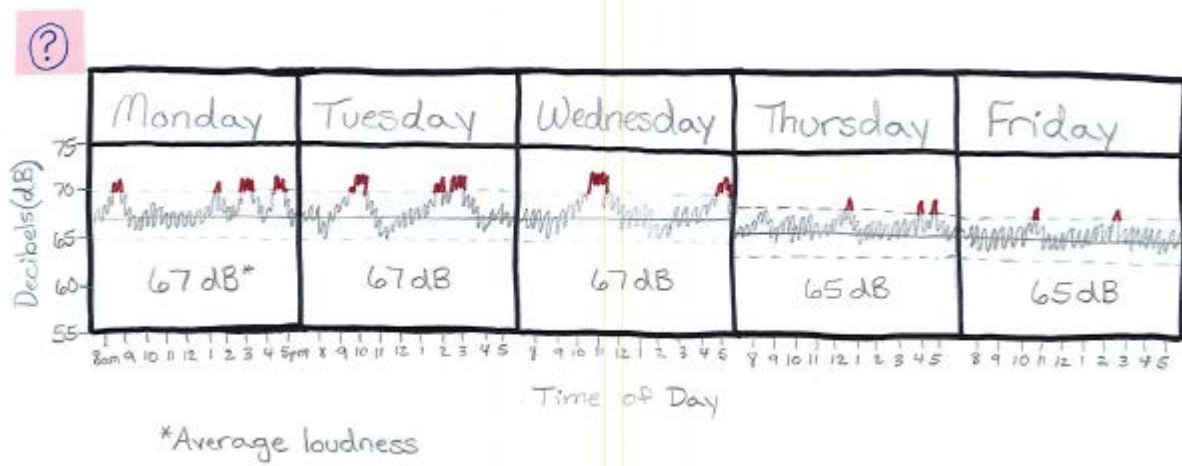
Figure 81: Iteration 2 displays – Loudness small multiple sparklines.
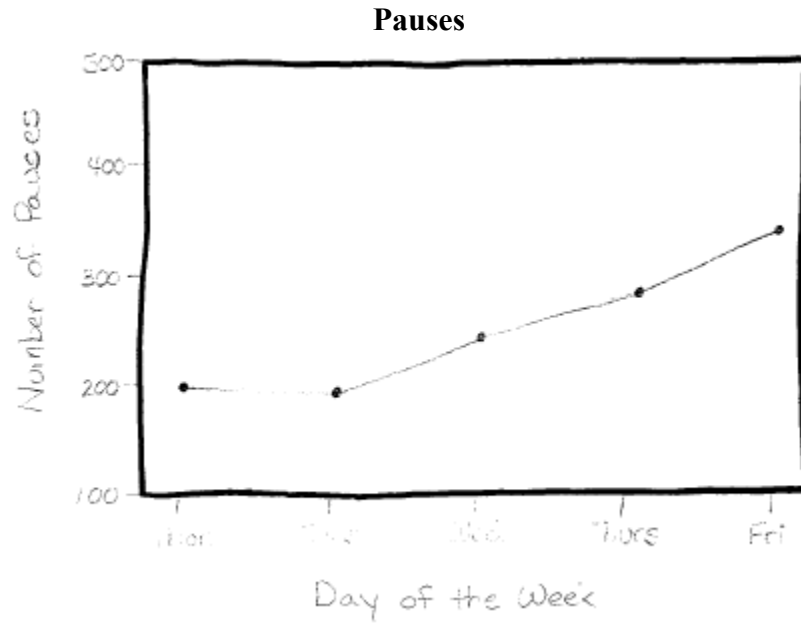
Pauses

Figure 82: Iteration 2 displays – Pause line graph.
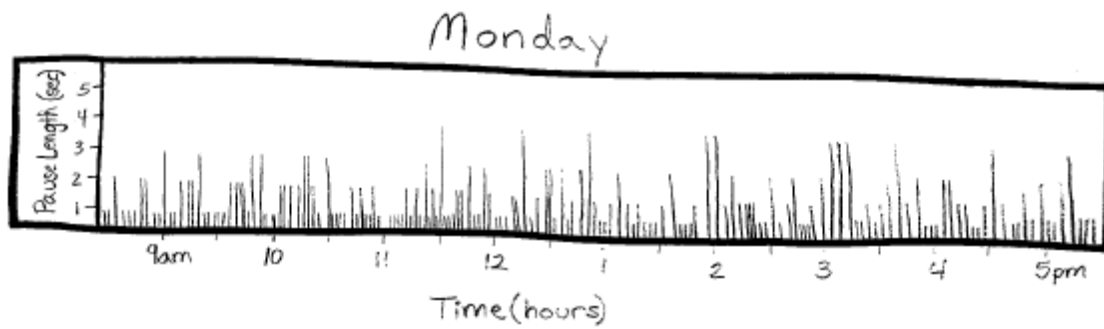
Figure 83: Iteration 2 displays - Individual pause time and length.
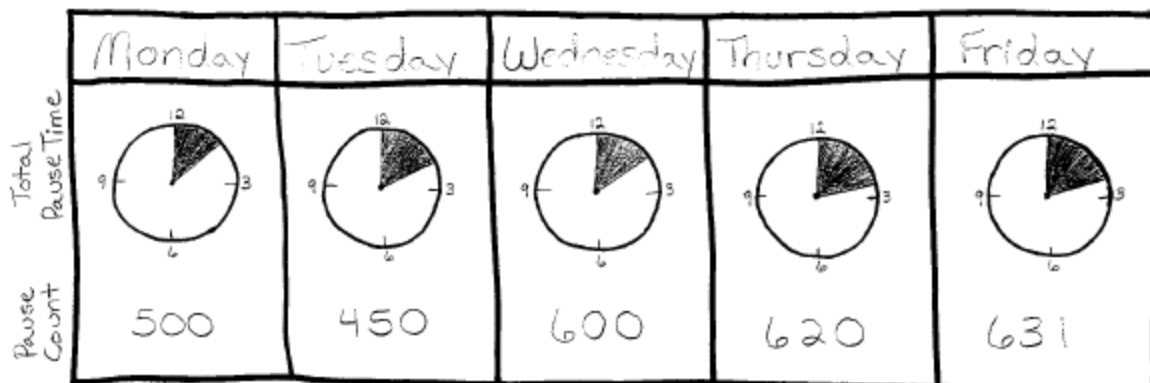
Figure 84: Iteration 2 displays – Pause clocks and counts.

169
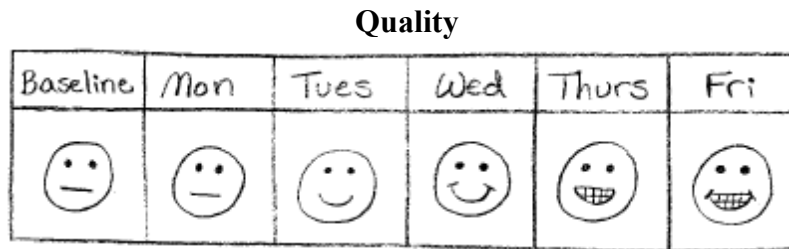
Quality

Figure 85: Iteration 2 displays – Quality small multiple smileys.
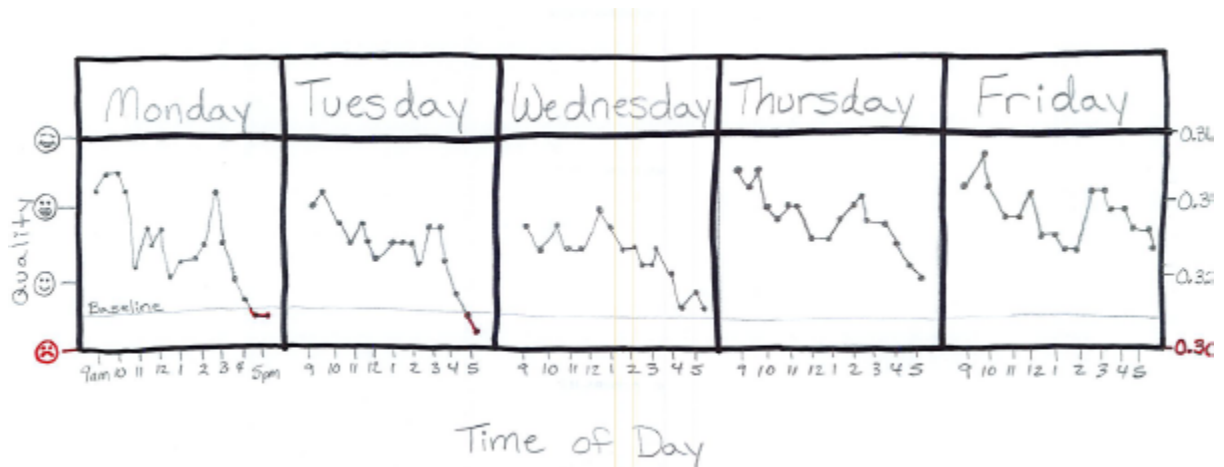
Figure 86: Iteration 2 displays – Quality small multiple line graphs.
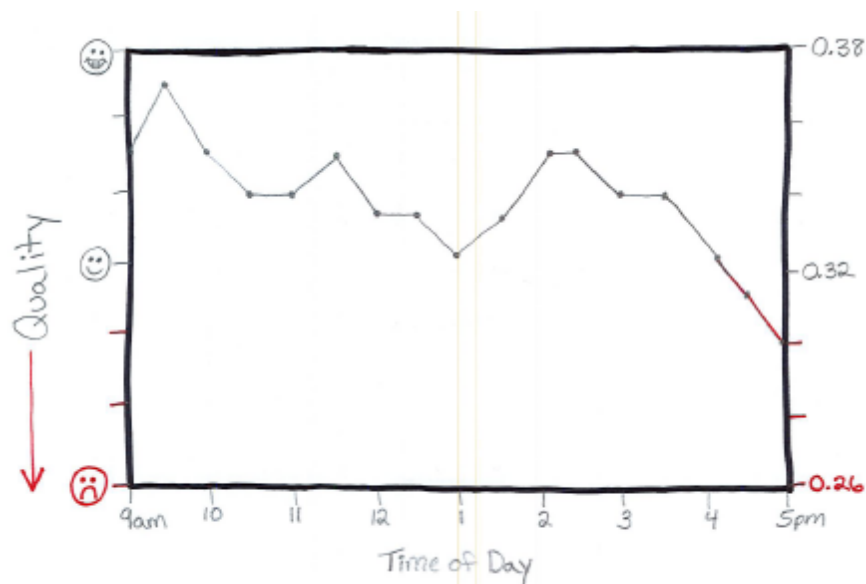
Figure 87: Iteration 2 displays – Quality single day line graph.
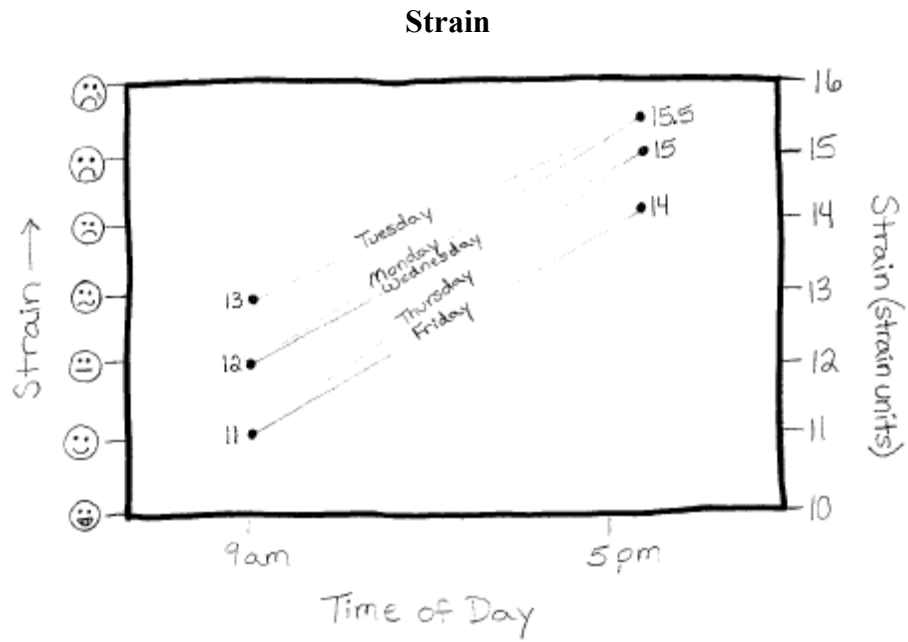
Strain

Figure 88: Iteration 2 displays - Strain labelled line graph.
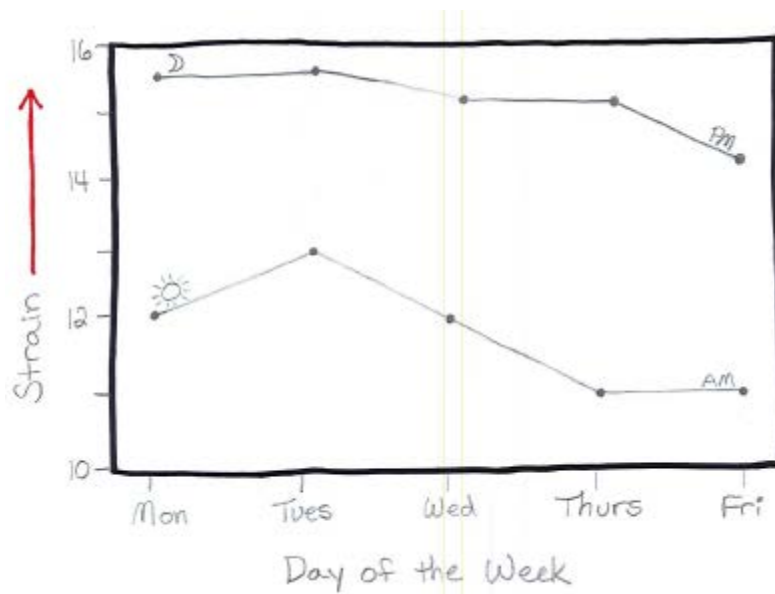
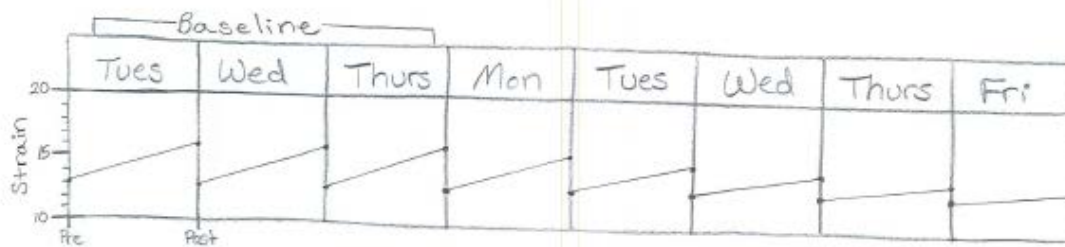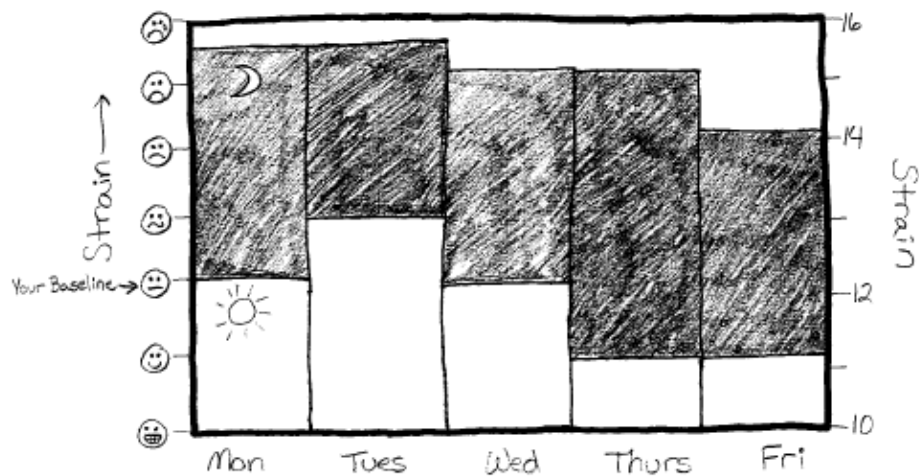Figure 89: Iteration 2 displays – Strain day and night line graph.

Multi-Measure
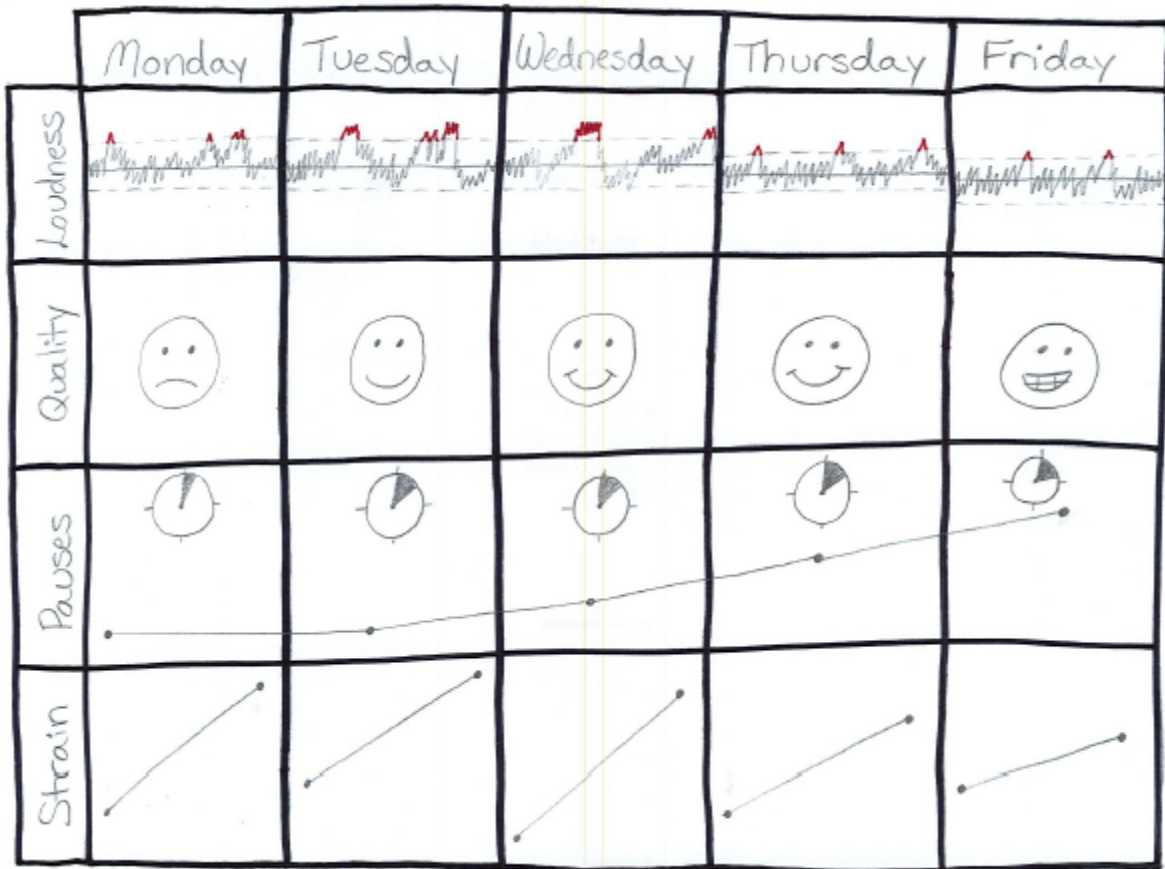
Figure 90: Iteration 2 displays – Multi-measure matrix.
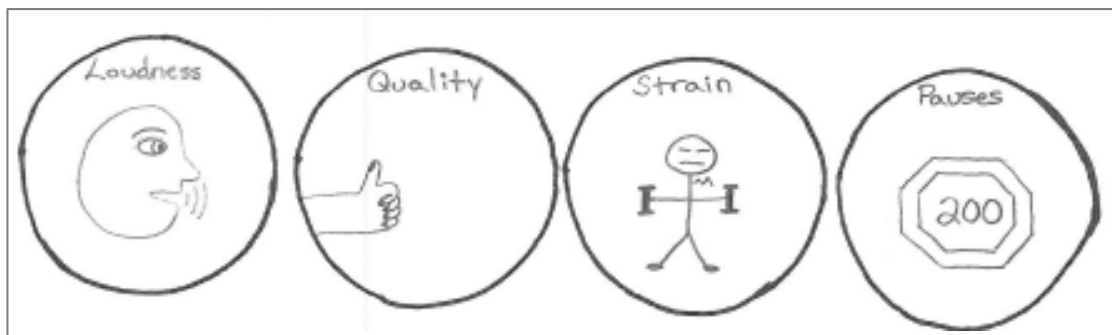
Figure 91: Iteration 2 displays - Dynamic icon array.

Multi-Measure
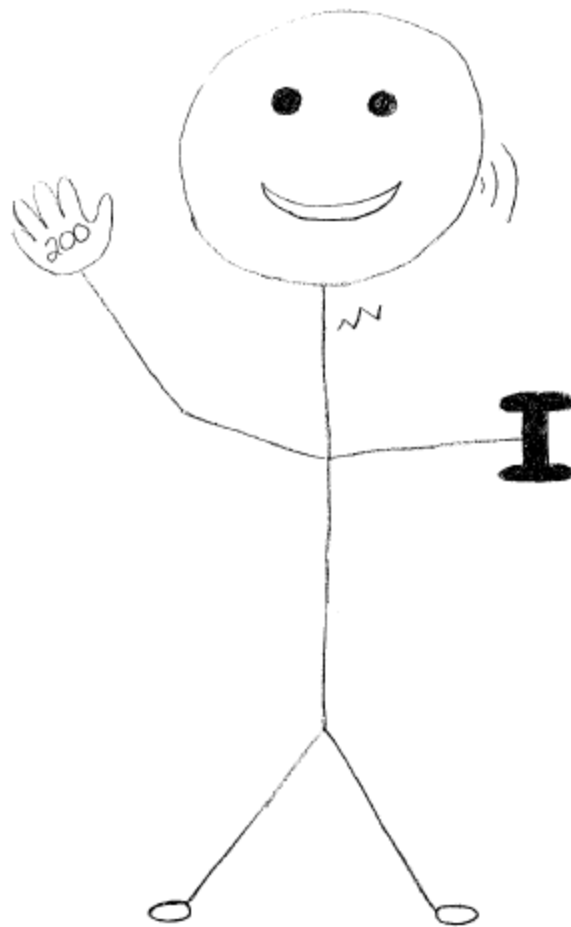
Figure 92: Iteration 2 displays – Multi-measure man.
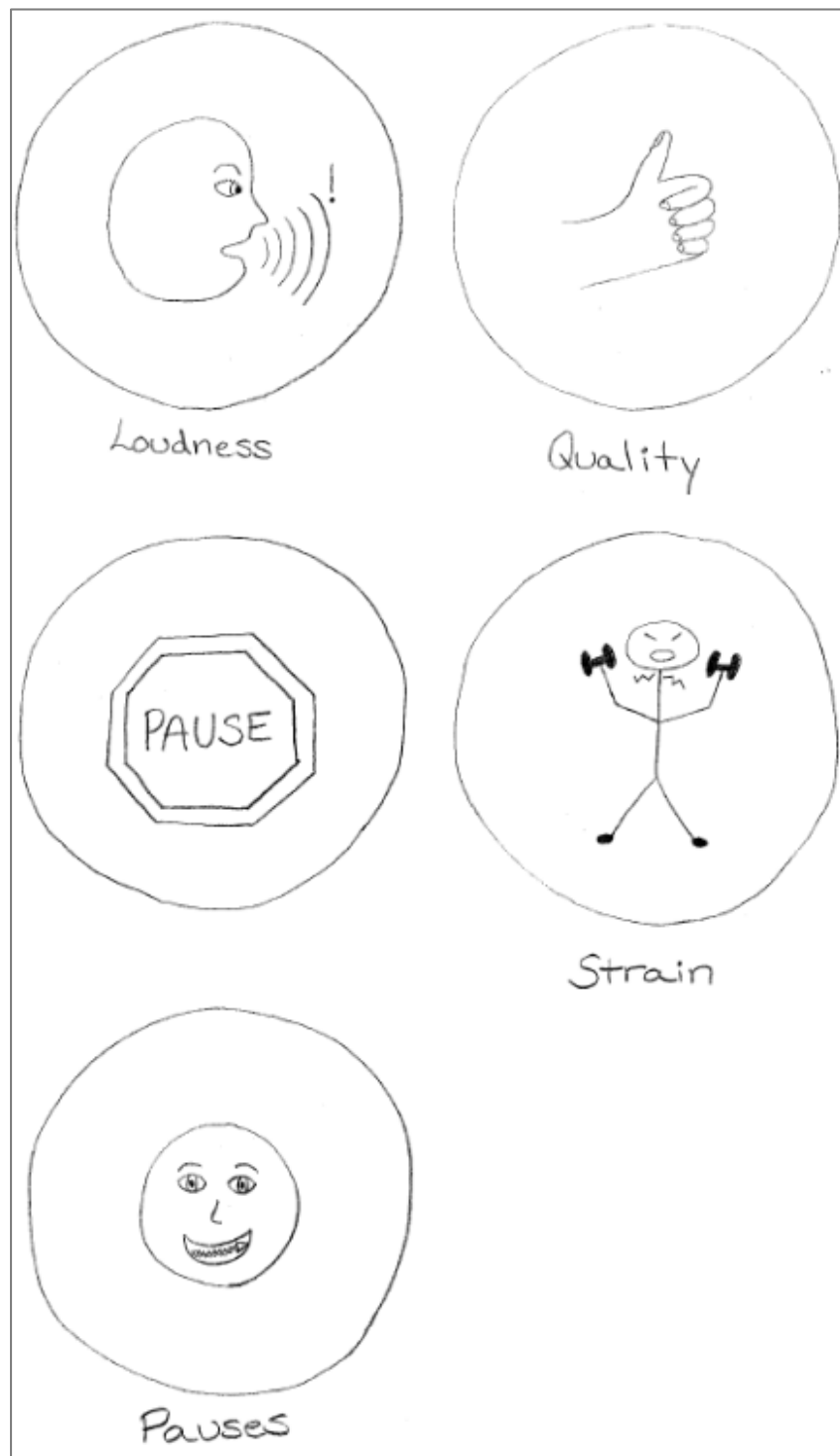
Icons

Figure 93: Iteration 3 displays –Icons for all four measures, with two pause options.

Loudness
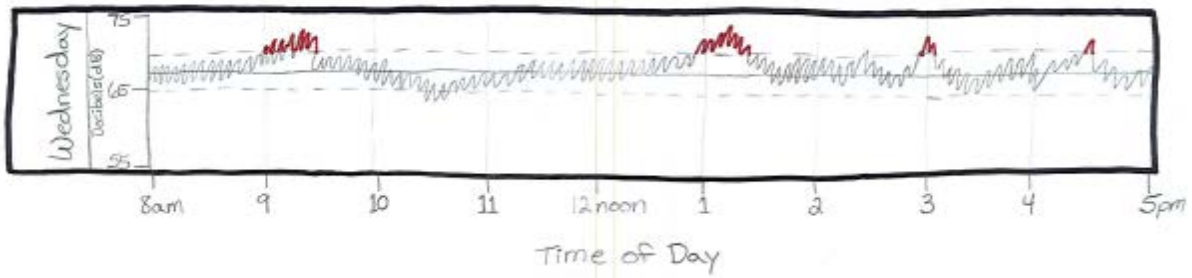
Figure 94: Iteration 3 displays –Loudness single day sparkline.

*Average loudness

Figure 95: Iteration 3 displays – Loudness small multiple sparklines.

Figure 96: Iteration 3 displays – Loudness single day sparkline with zeros.

Pauses
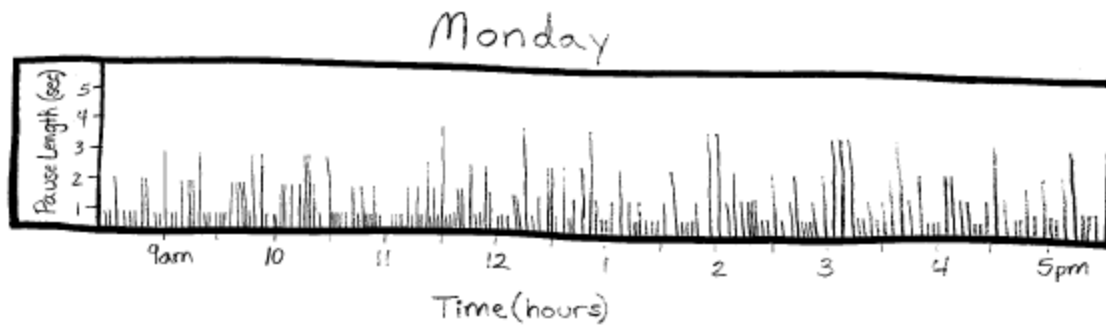
Figure 97: Iteration 3 displays – Individual pause time and length.
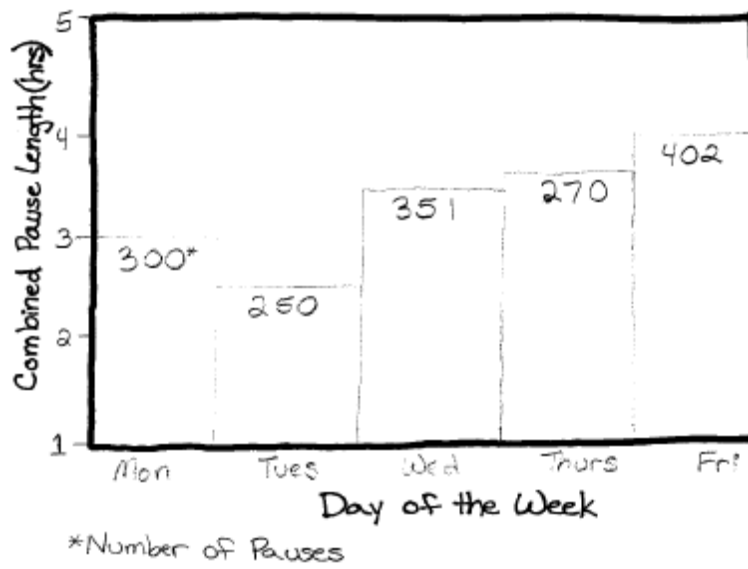
Figure 98: Iteration 3 displays – Pause bar graph with counts.

Quality
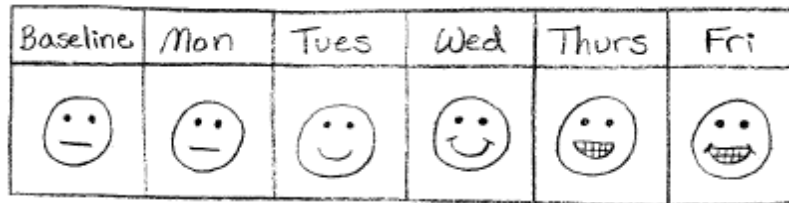
Figure 99: Iteration 3 displays – Quality small multiple smileys.
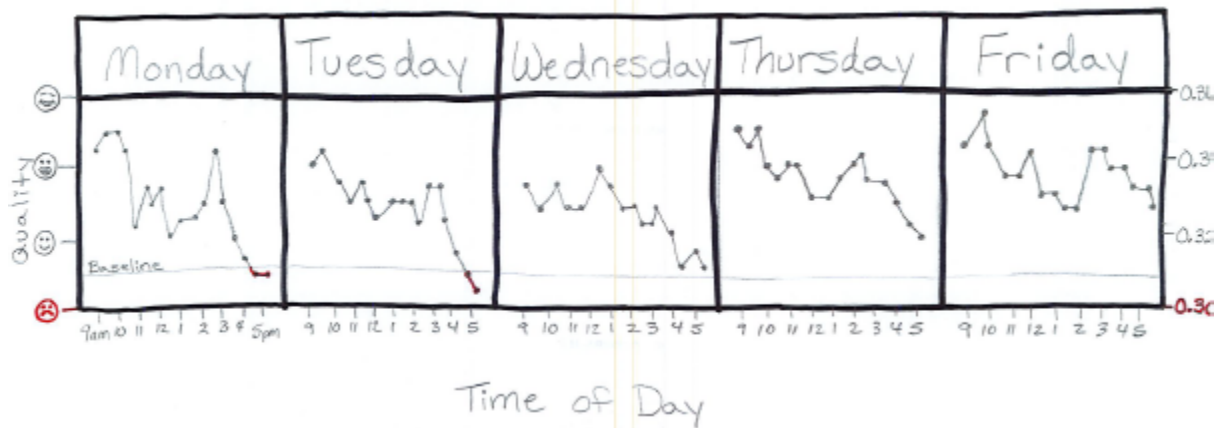
Figure 100: Iteration 3 displays – Quality small multiple line graphs.
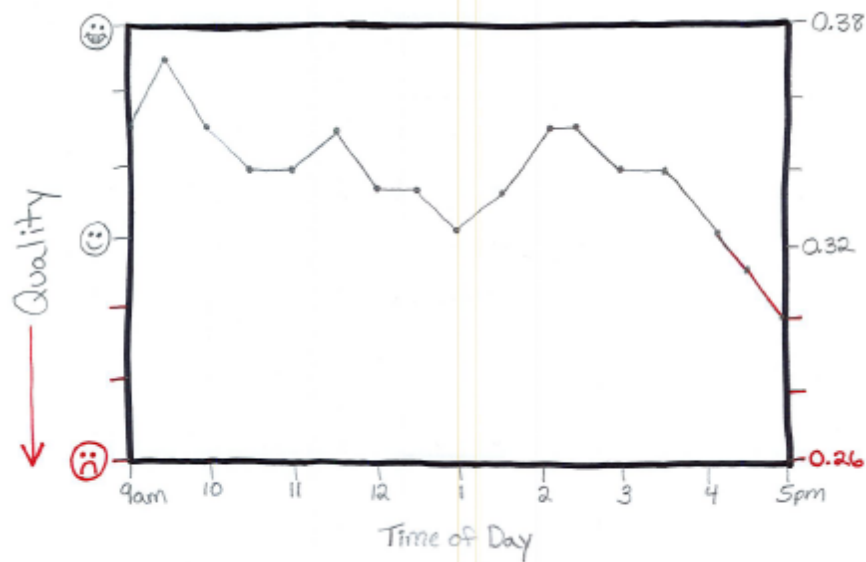
Figure 101: Iteration 3 displays – Quality single day line graph.
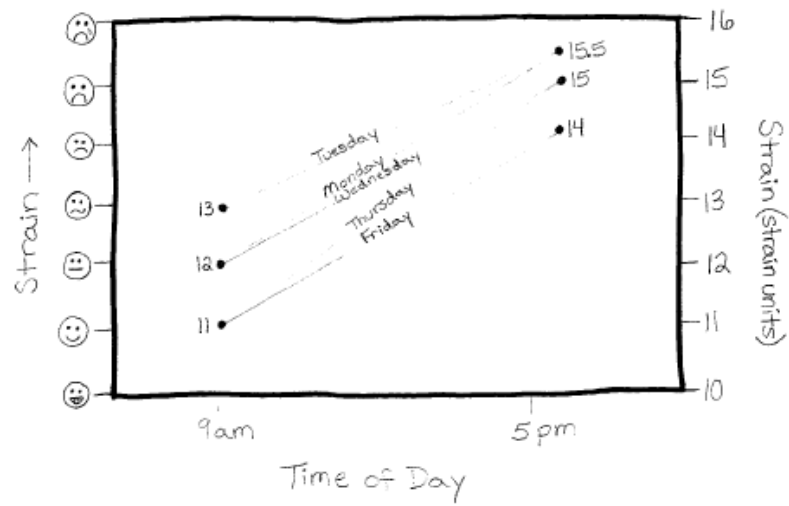
Strain

Figure 102: Iteration 3 displays – Strain labelled line graph.

Figure 103: Iteration 3 displays – Strain small multiple line graph.

Figure 104: Iteration 3 displays – Strain day and night bar graph.

Multi-Measure
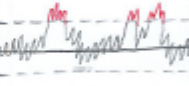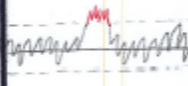
	Mon	Tues	Wed	Thurs	Fri
Quality	☹	☺	☺	☺	😁
Pauses	500	450	600	620	631
Strain	15	15	14.5	14.5	14
Loudness					

Figure 105: Iteration 3 displays – Multi-measure matrix.
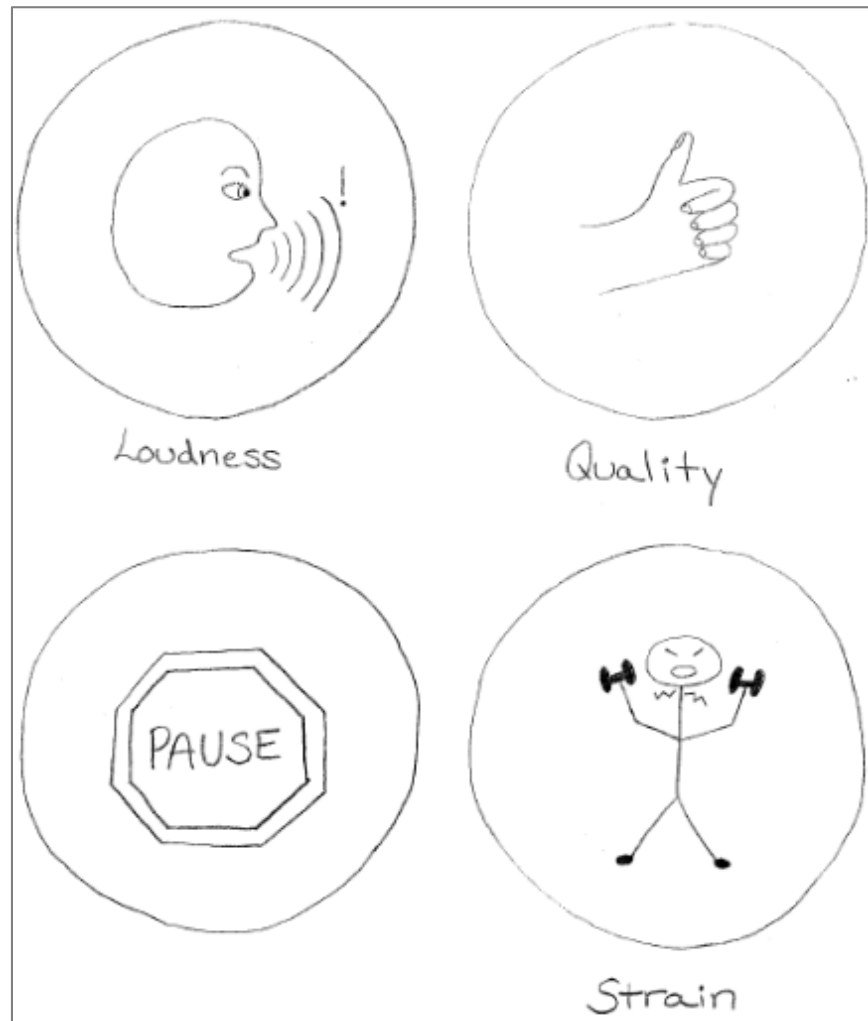
Appendix H: Iteration 4 Feedback Displays:

Icons

Figure 106: Iteration 4 displays – Icons for the four measures.

Loudness

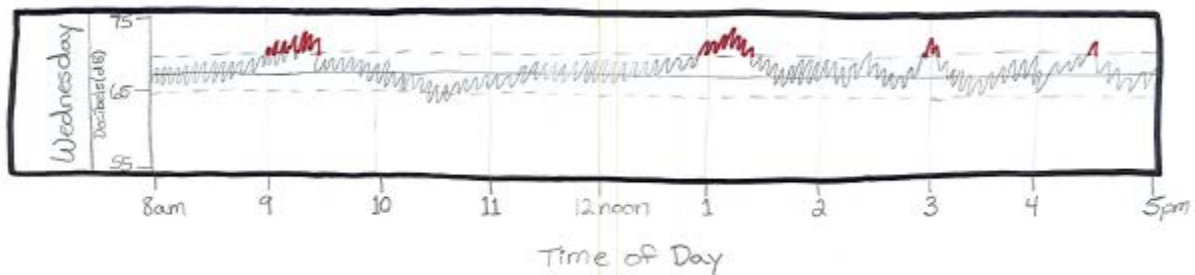Figure 107: Iteration 4 displays – Loudness multi-day danger zone counts.

Figure 108: Iteration 4 displays – Loudness single day sparkline.
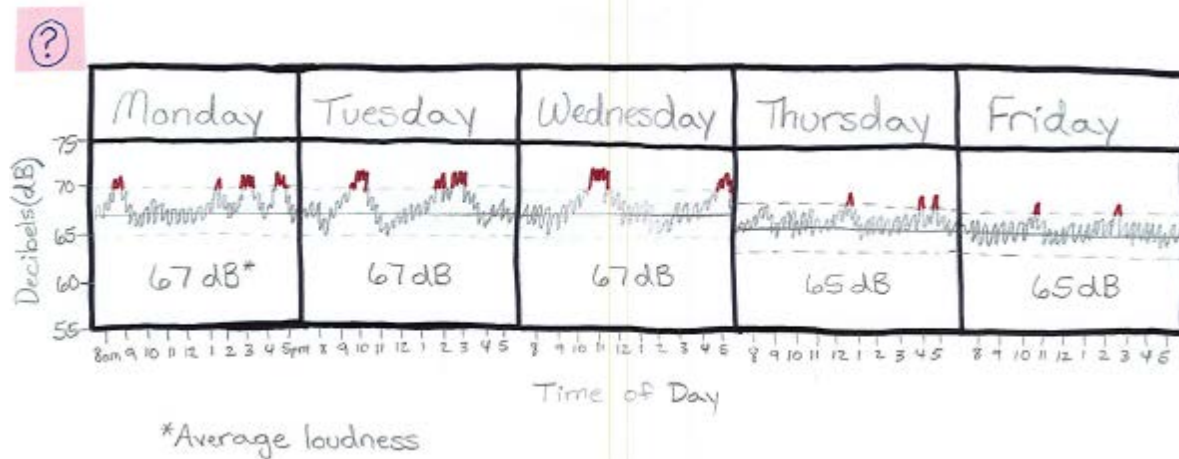
Figure 109: Iteration 4 displays – Loudness small multiple sparklines.

Pauses

Figure 110: Iteration 4 displays – Pause multi-day counts.
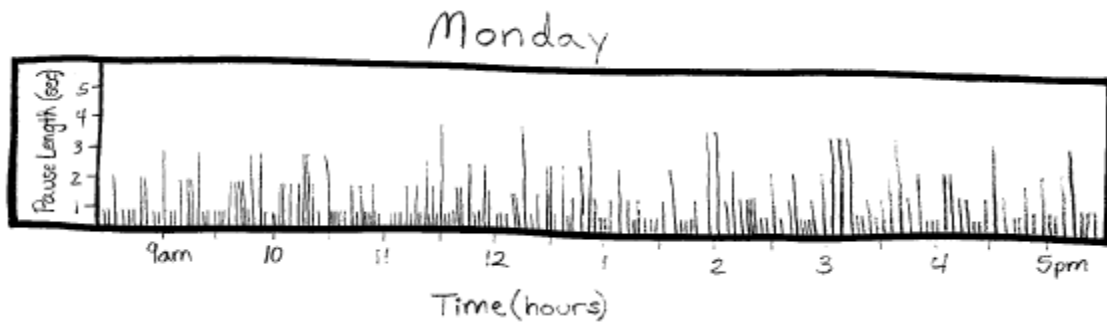
Figure 111: Iteration 4 displays – Individual pause time and length.

Figure 112: Iteration 4 displays – Pause small multiple bar graphs.

Pauses

Figure 113: Iteration 4 displays – Pause bar graph.

Quality

Figure 114: Iteration 4 displays – Quality small multiple smileys.
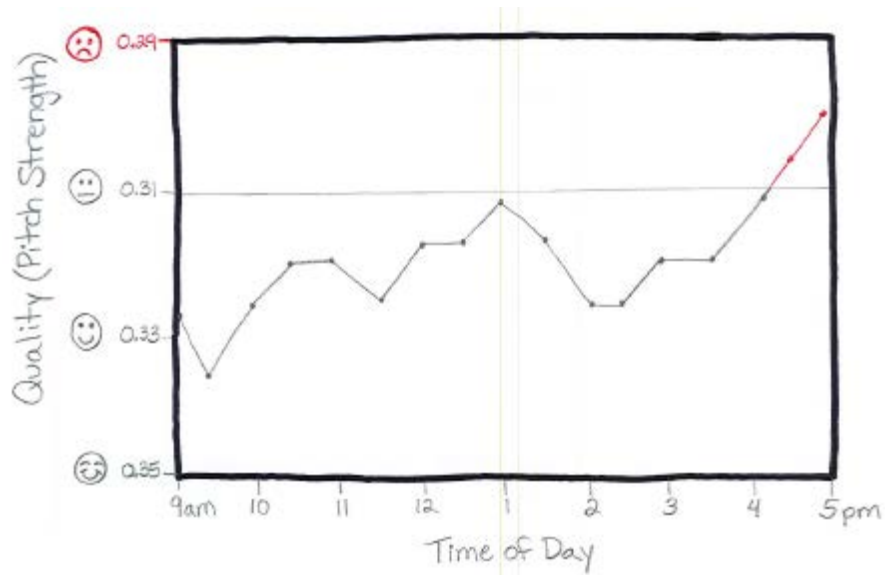
Figure 115: Iteration 4 displays – Quality single day line graph.
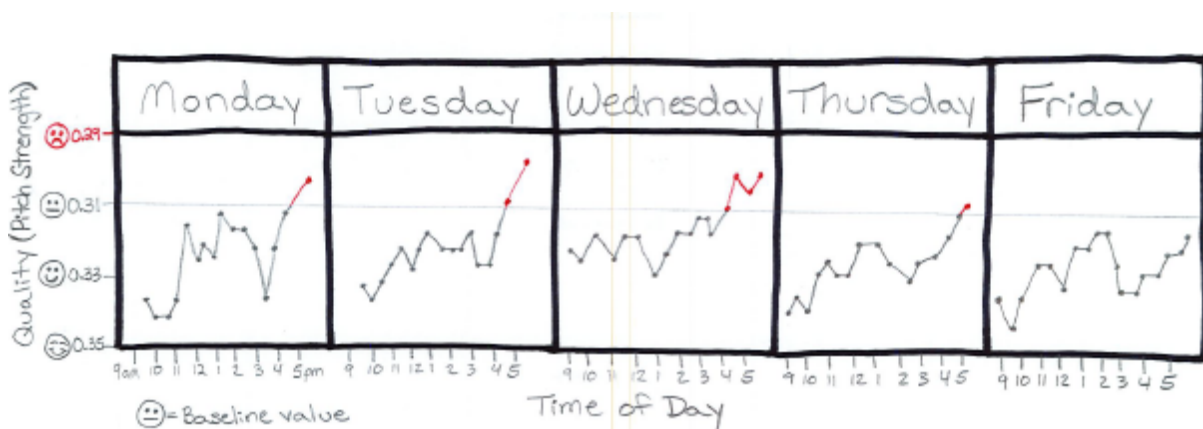
Figure 116: Iteration 4 displays – Quality small multiple line graphs.

Strain

*Compared with baseline

Figure 117: Iteration 4 displays – Strain small multiple arrows.
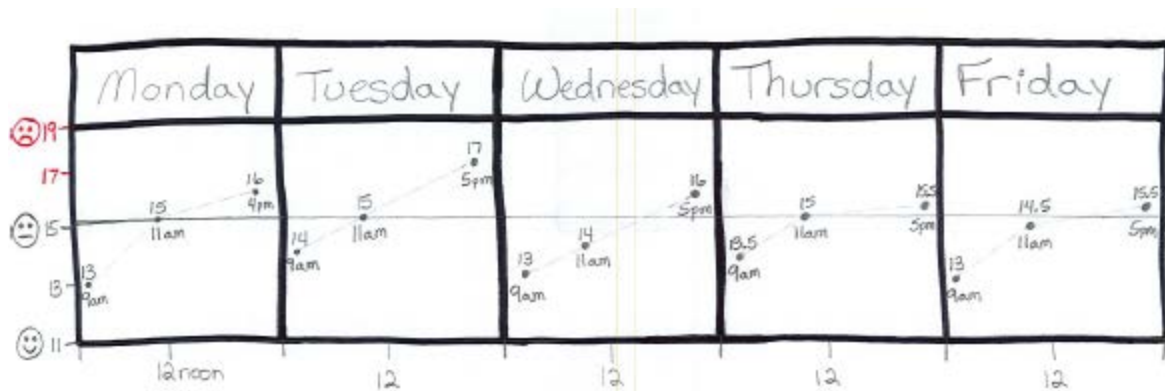
Figure 118: Iteration 4 displays – Strain small multiple time points.
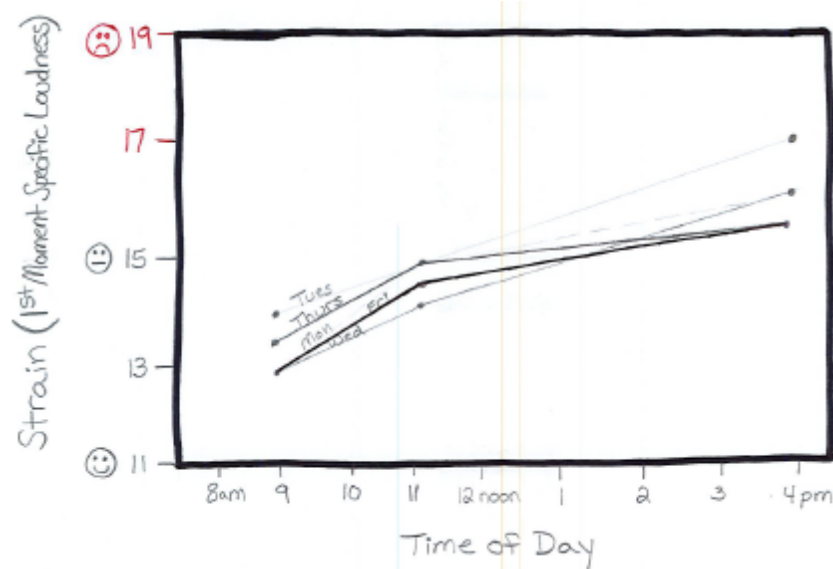
Figure 119: Iteration 4 displays – Strain labelled line graphs.

Multi-Measure

Figure 120: Iteration 4 displays – Multi-measure matrix.

Appendix I: Final Feedback Displays
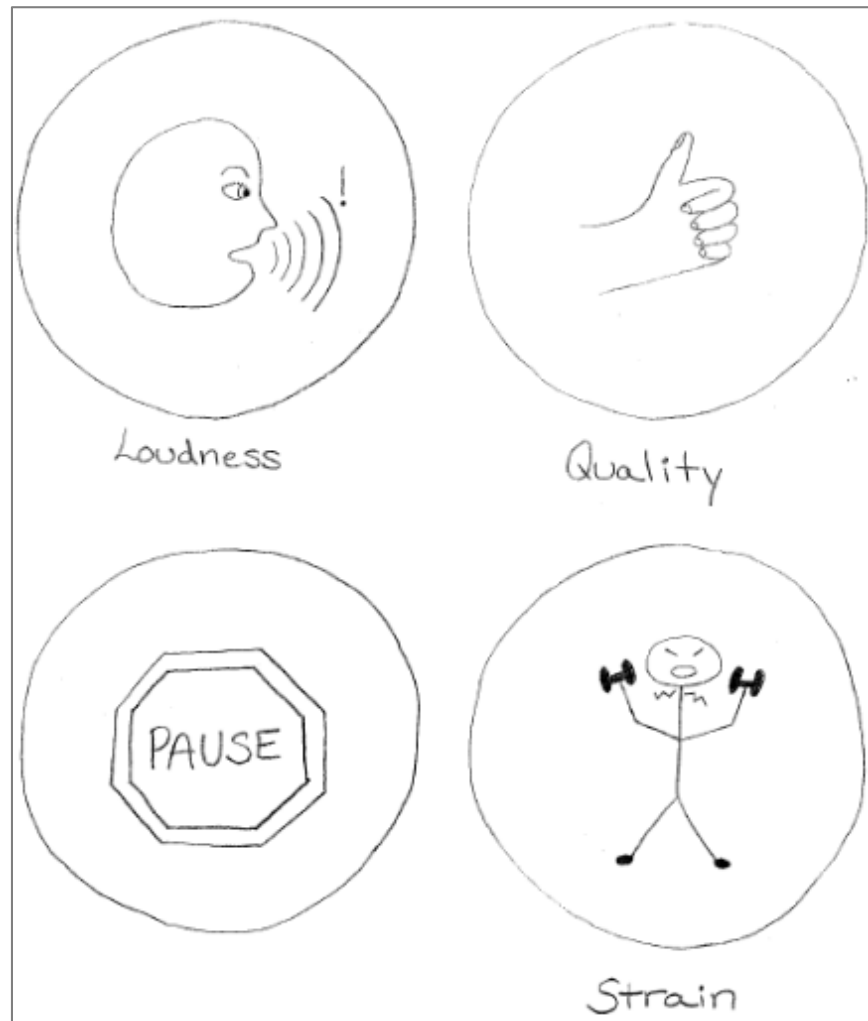
Icons

Figure 121: Final displays – Icons for the four measures.
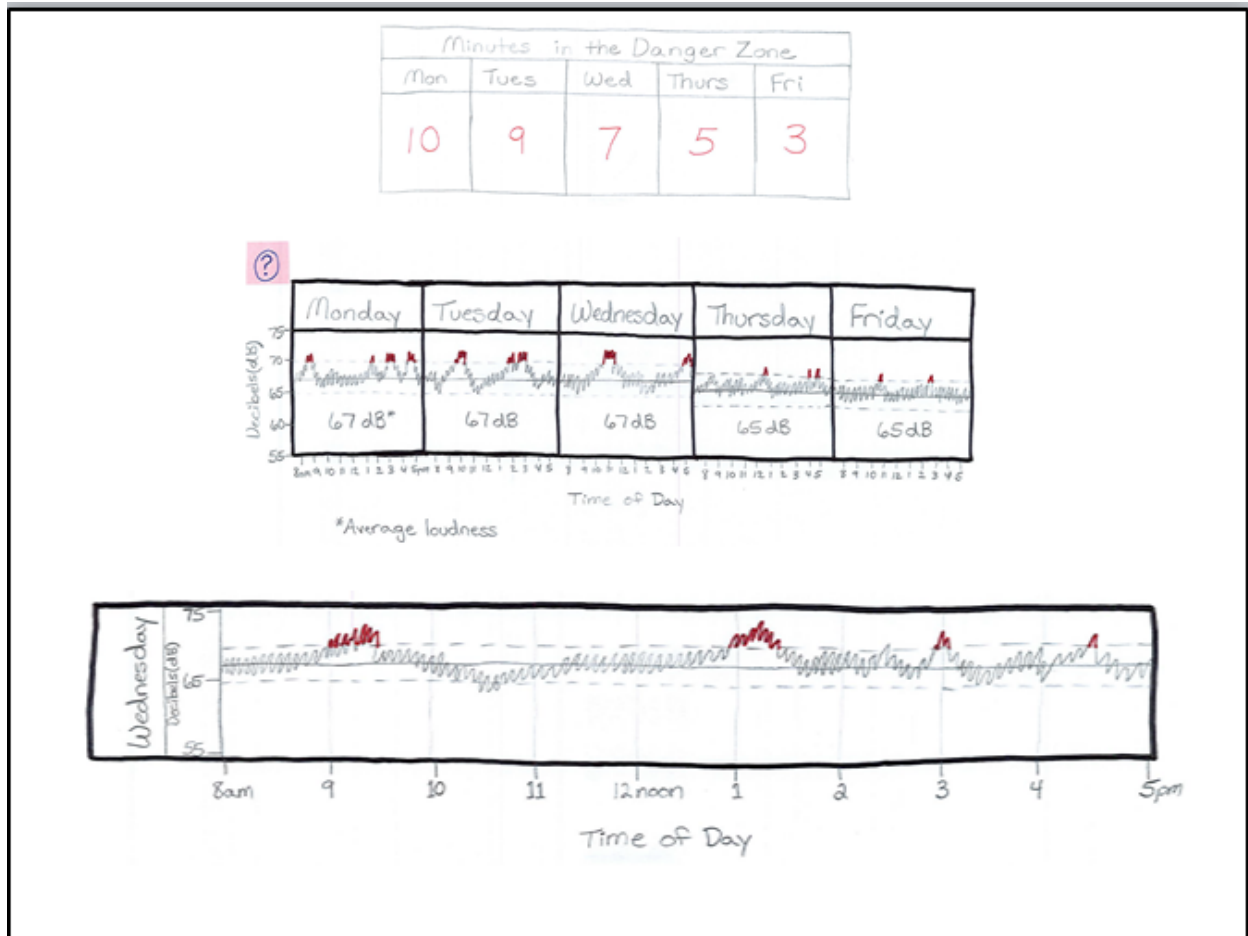
Loudness

Figure 122: Final displays – Layered structure for loudness.

Pauses
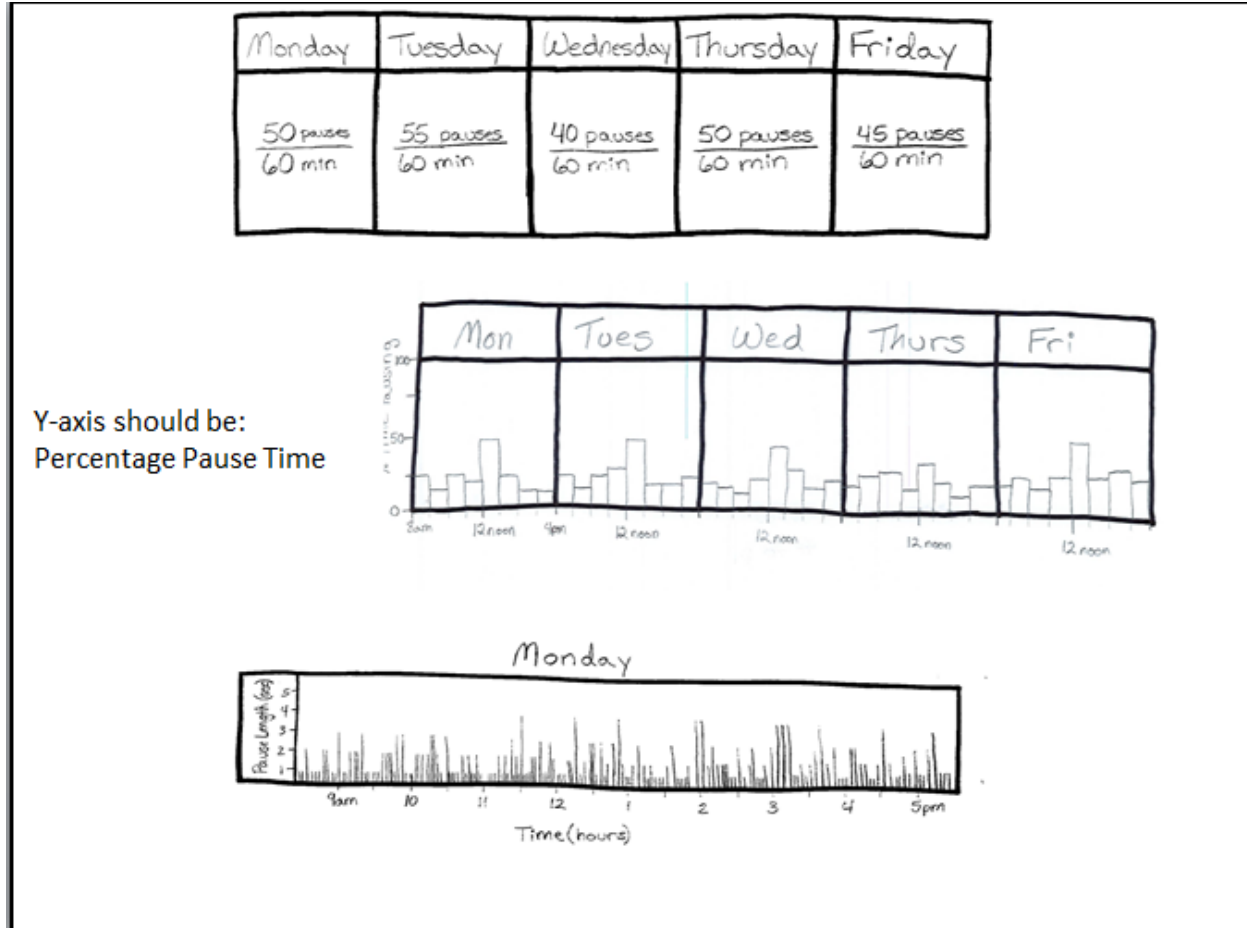
Figure 123: Final displays – Layered structure for pauses.

Quality
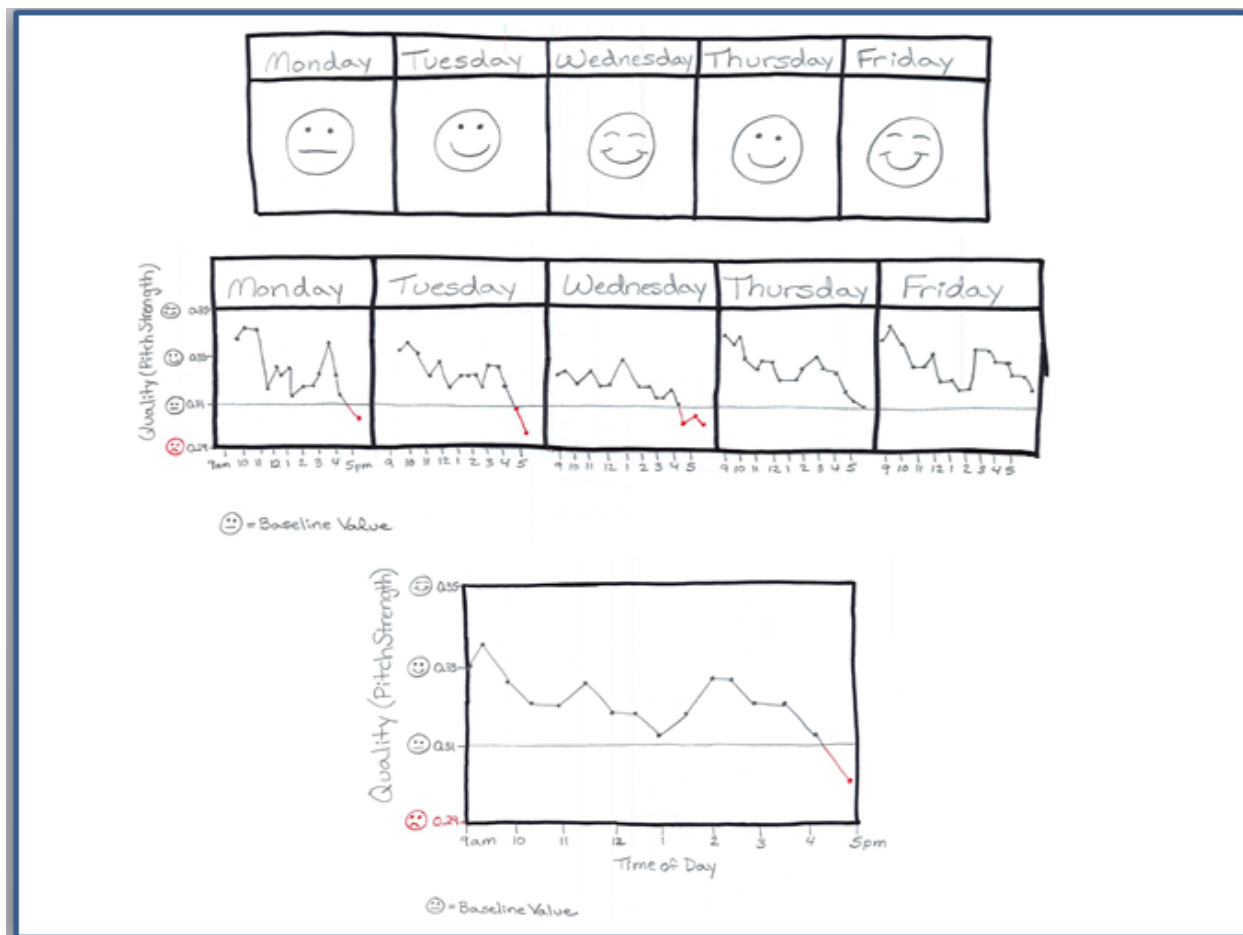
Figure 124: Final displays – Layered structure for quality.
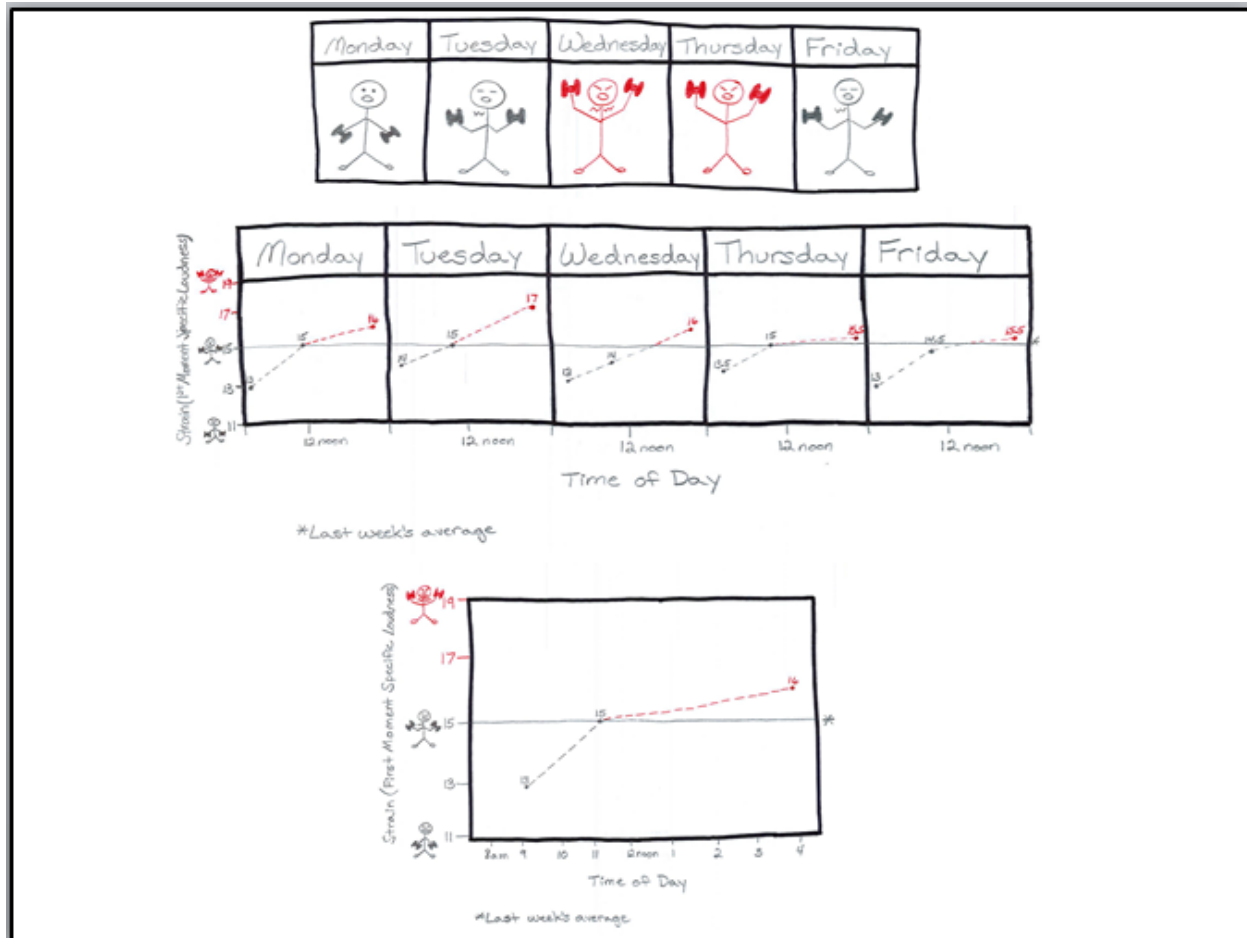
Figure 125: Final displays – Layered structure for strain.

Appendix J: Part 2 Midpoint and Final Semi-Structured Interview Questions

Midpoint Semi-Structured Interview

Overall
1. What is your overall impression of the study so far?
2. What did you like most?
    a. Why?
3. What did you like least?
    a. Why?

Wearing the Device (questions based on Hunter, 2012))
1. What did you think about wearing the collar and recorder?
2. Did you experience any discomfort with the equipment?
    a. How did you resolve it?
3. Did you experience any difficulties with the equipment?
    a. How did you resolve them?
4. Do you have any suggestions for how to improve the recording system?

Strategies
1. Did you change how you talked during the study?
    a. How?
2. Did you change how much you talked during the study?
    a. How?
3. Did you change how loud you talked during the study?
    a. How?

How much did you change <u>how</u> you talked? Indicate (make a vertical tick mark) on the scale below:

_____

Not at All   Extremely

How much did you change <u>how much</u> you talked? Indicate (make a vertical tick mark) on the scale below:

_____

Not at All   Extremely

How much did you change <u>how loud</u> you talked? Indicate (make a vertical tick mark) on the scale below:

_____

Not at All   Extremely

Final Semi-Structured Interview

Overall
1. What was your overall impression of the study?
2. What did you like most?
    a. Why?
3. What did you like least?
    a. Why?

Strategies
1. Did you change how you talked during the second part of the study?
    a. How?
2. Did you change how much you talked during the second part of the study?
    a. How?
3. Did you change how loud you talked during the second part of the study?
    a. How?

Feedback
1. What did you think about the feedback?
2. Do you feel like the feedback was helpful?
3. What did you think was most helpful?
    a. Why?
    b. How did it influence your talking?
4. What did you think was least helpful?
    a. Why?
5. What additional feedback would be helpful?
    a. Why?
6. What did you think about how the feedback was presented?
7. How would you improve the way the feedback is presented? I have a pen and paper available if you'd like to draw an example.

Sharing

1. Did you share your feedback information with anyone?

    a. Who?

    b. Why?

    c. What were their thoughts?

    d. (If shared information with other participants) Did you find this helpful?

Recommendations

1. If a similar type of system were available for personal use, what should it include?
    a. Any other features you would add?
2. One future idea is adding a social component, where people can share their data with others. What are your thoughts?
3. How much do you think people would be willing to pay for this kind of system for personal use? Why?
4. If employers made these systems available for employee use, do you think employees would use them? Why?

How much did you change <u>how</u> you talked? Indicate (make a vertical tick mark) on the scale below:

_____
Not at All   Extremely

How much did you change <u>how much</u> you talked? Indicate (make a vertical tick mark) on the scale below:

_____
Not at All   Extremely

How much did you change <u>how loud</u> you talked? Indicate (make a vertical tick mark) on the scale below:

_____
Not at All   Extremely

Appendix K: Sample Feedback (Part 2.1)
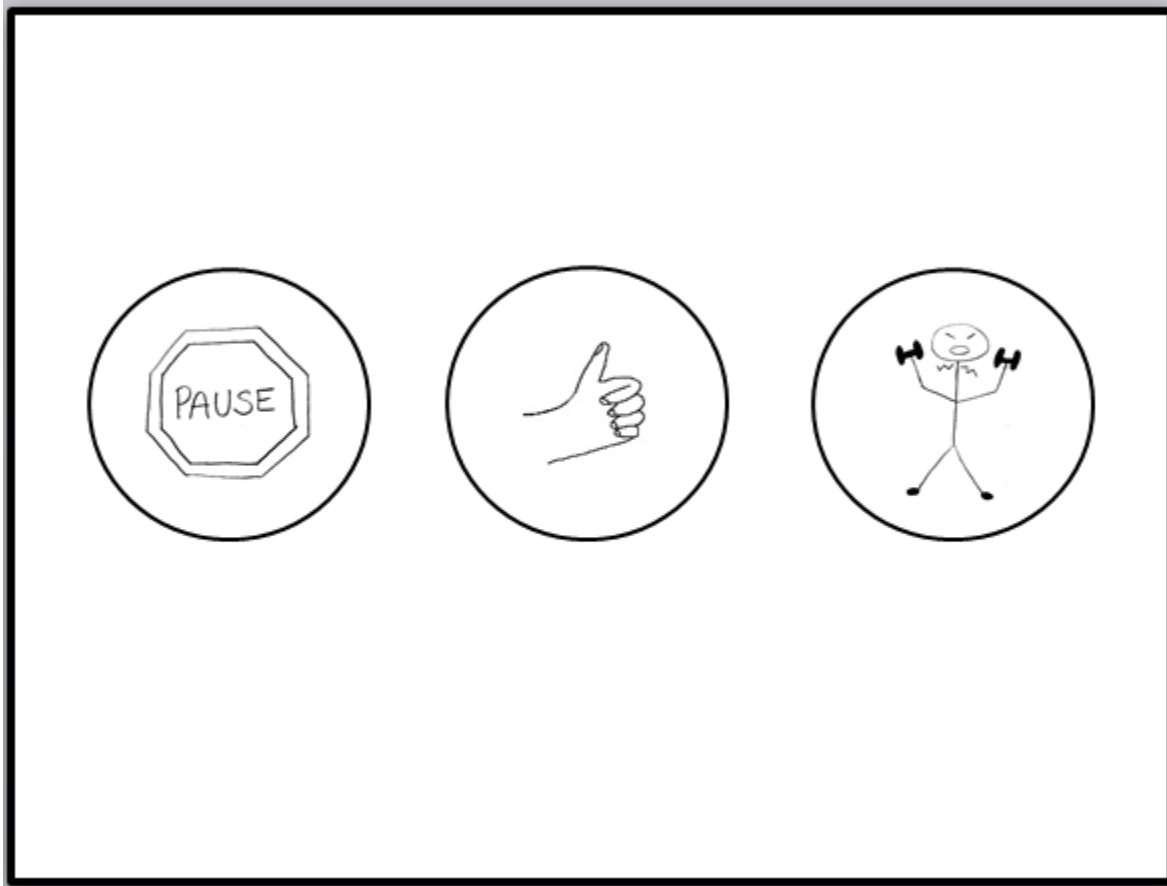
Figure 126: Initial image seen by participants. Icons are displayed in a different random order for each participant. This participant's order is: pauses, quality, and strain.

Day 3	Day 4	Day 5	Day 6	Day 7
$\dfrac{24\ Pauses}{15\ Minutes}$	$\dfrac{44\ Pauses}{15\ Minutes}$	$\dfrac{31\ Pauses}{15\ Minutes}$	$\dfrac{91\ Pauses}{15\ Minutes}$	$\dfrac{38\ Pauses}{15\ Minutes}$

Figure 127: First pause display. This display shows the pause count (number of pauses equal to or greater than one second in length) over the course of the reading VLT.
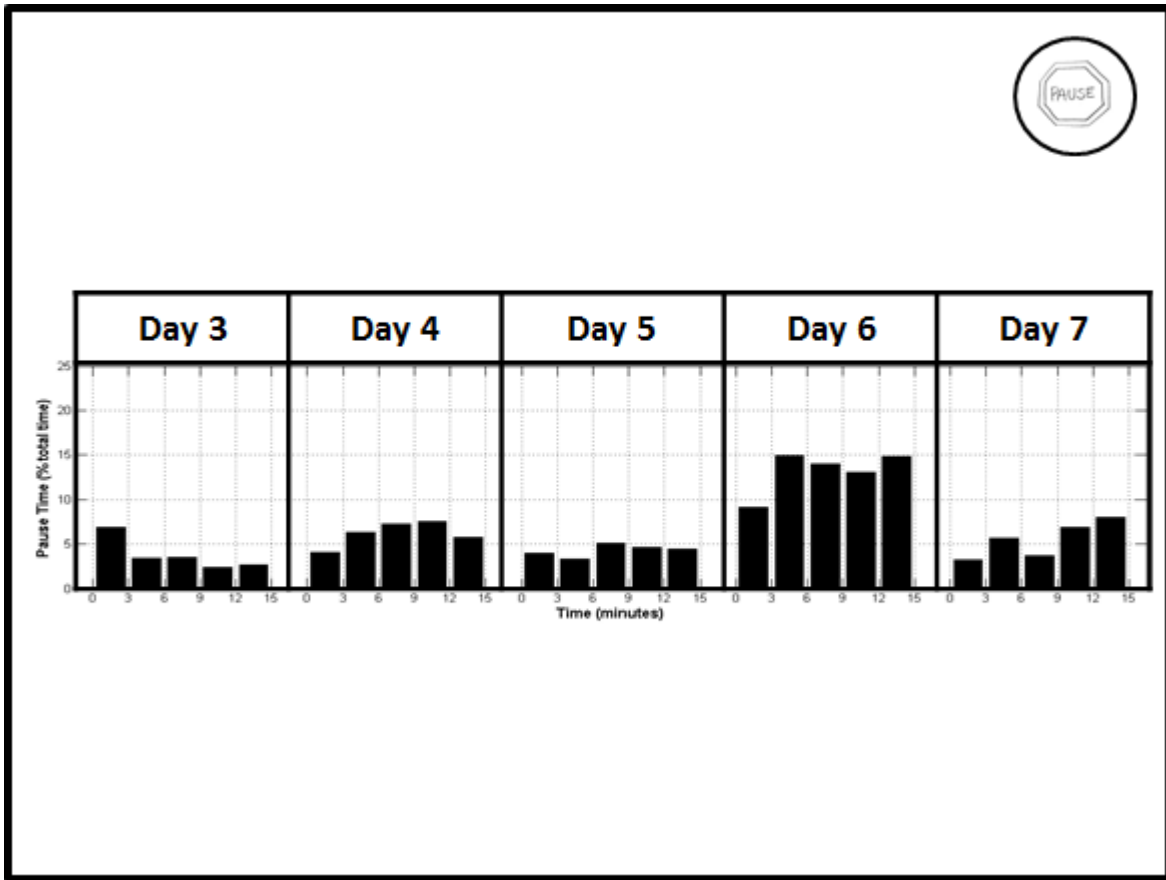
Figure 128: Second pause display. This display shows the amount of time (% total time) spent in pauses equal to or greater than one second in length for each 3 minutes of reading.
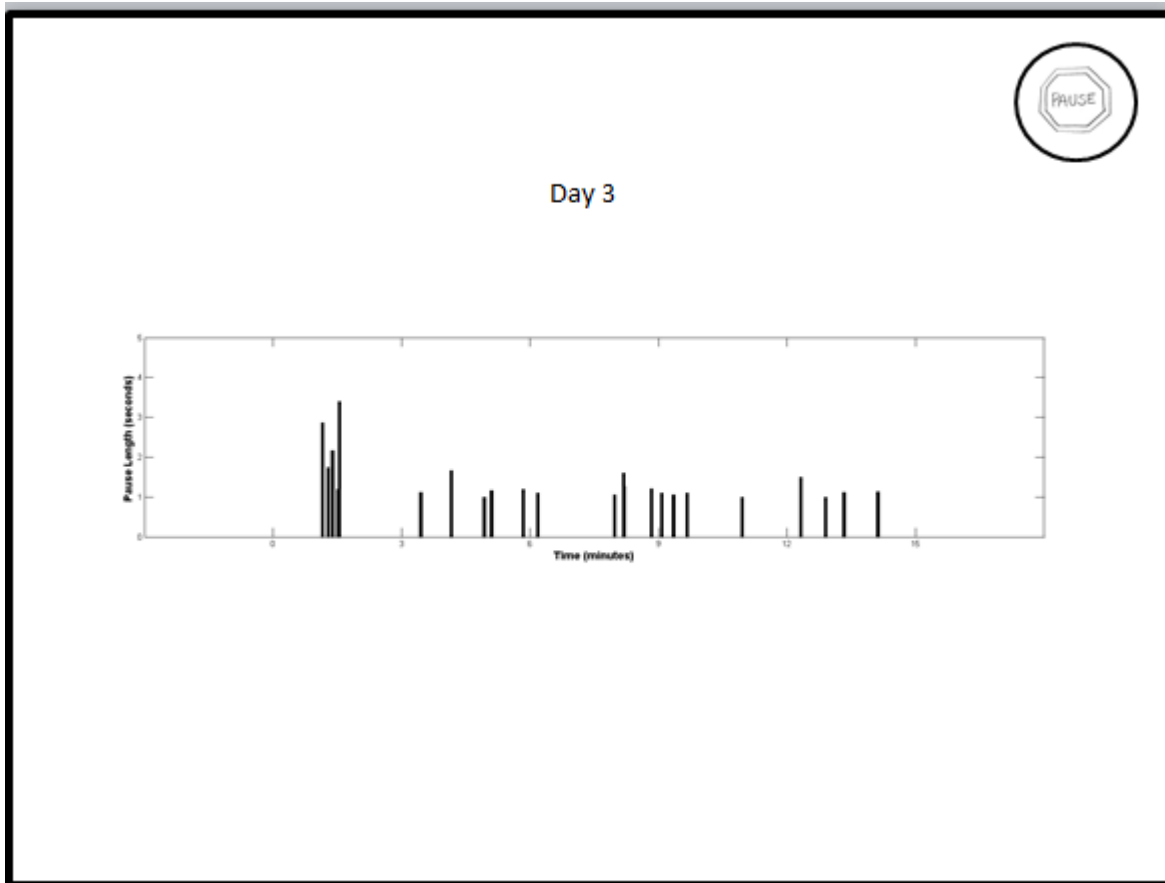
Figure 129: Third pause display. Participants had one of these for each day. This display shows when these pauses of a second or greater occurred, and their individual durations.
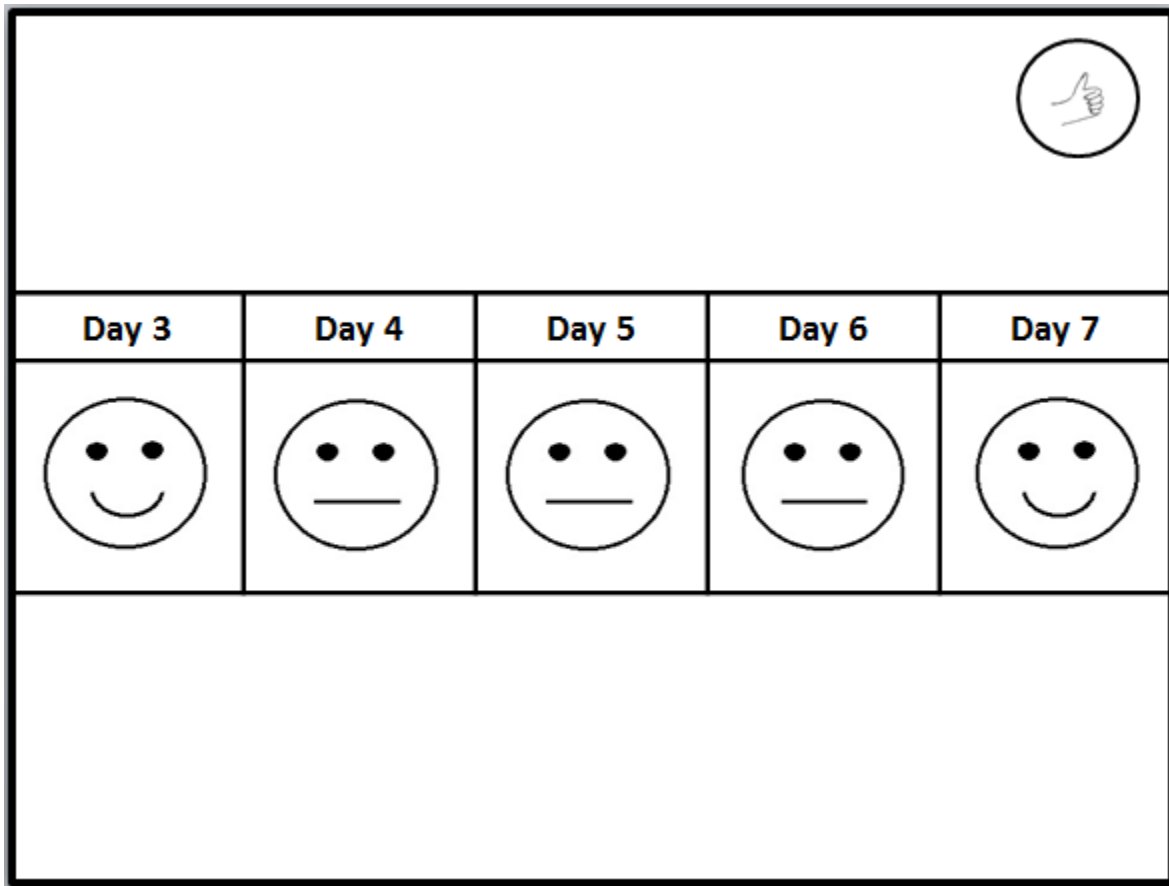
Figure 130: First quality display. The smileys indicate the average value for quality for each day (across all 15 minutes of reading). Note that the smiley with a straight line for a mouth is equal to "baseline."
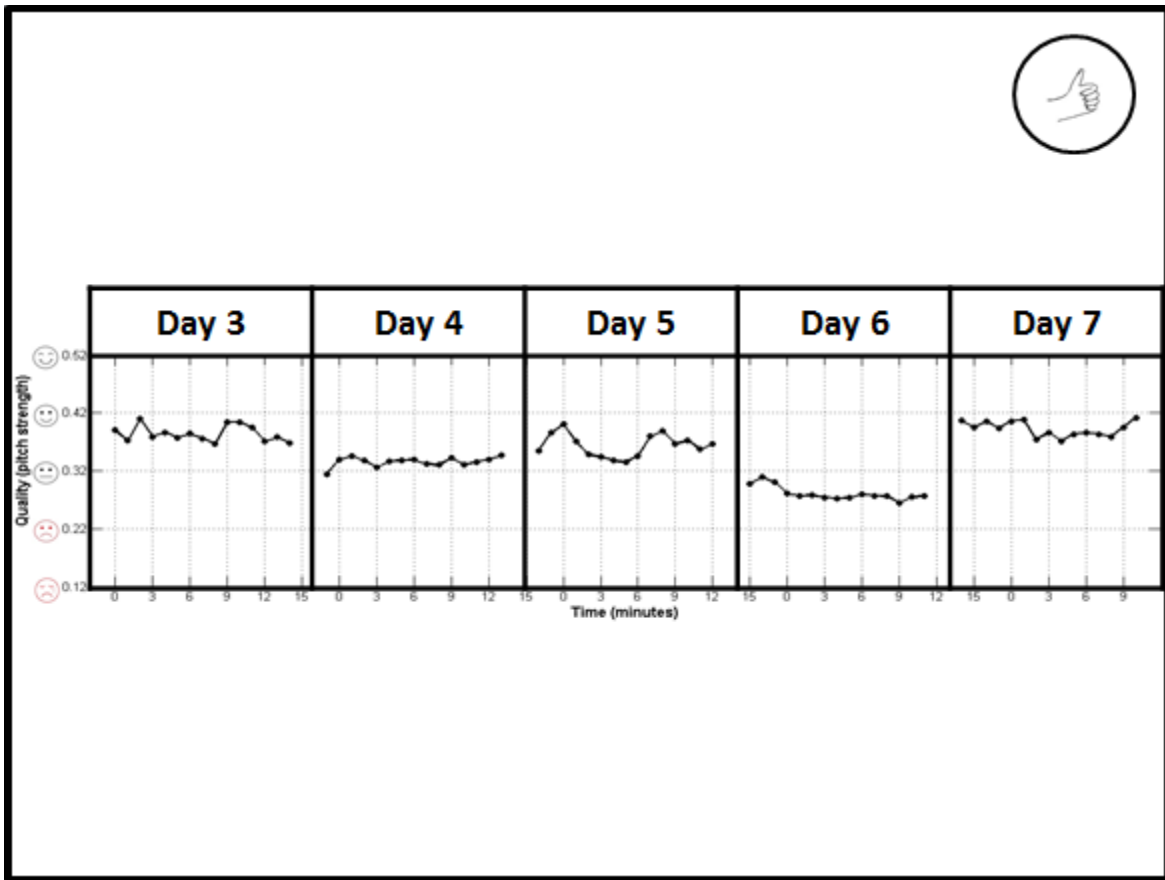
Figure 131: Second quality display. The "baseline" smiley is the average of the first minute of the baseline recordings (3 total).

Only the first minute was analyzed to ensure that there were no vocal fatigue effects. The range around the baseline was based on the range of pitch strength values seen in voice signal typing (Kopf, Shrivastav, Eddins, Skowronski, & Hunter, 2014), where the average value for Type 1 voices was 0.45 and the average value for Type 3 voices was 0.05. This range was chosen to ensure that all values for a given participant would fall within that range (e.g., even if fatigue leads to a lessening of voice quality, it should still fall within this range).
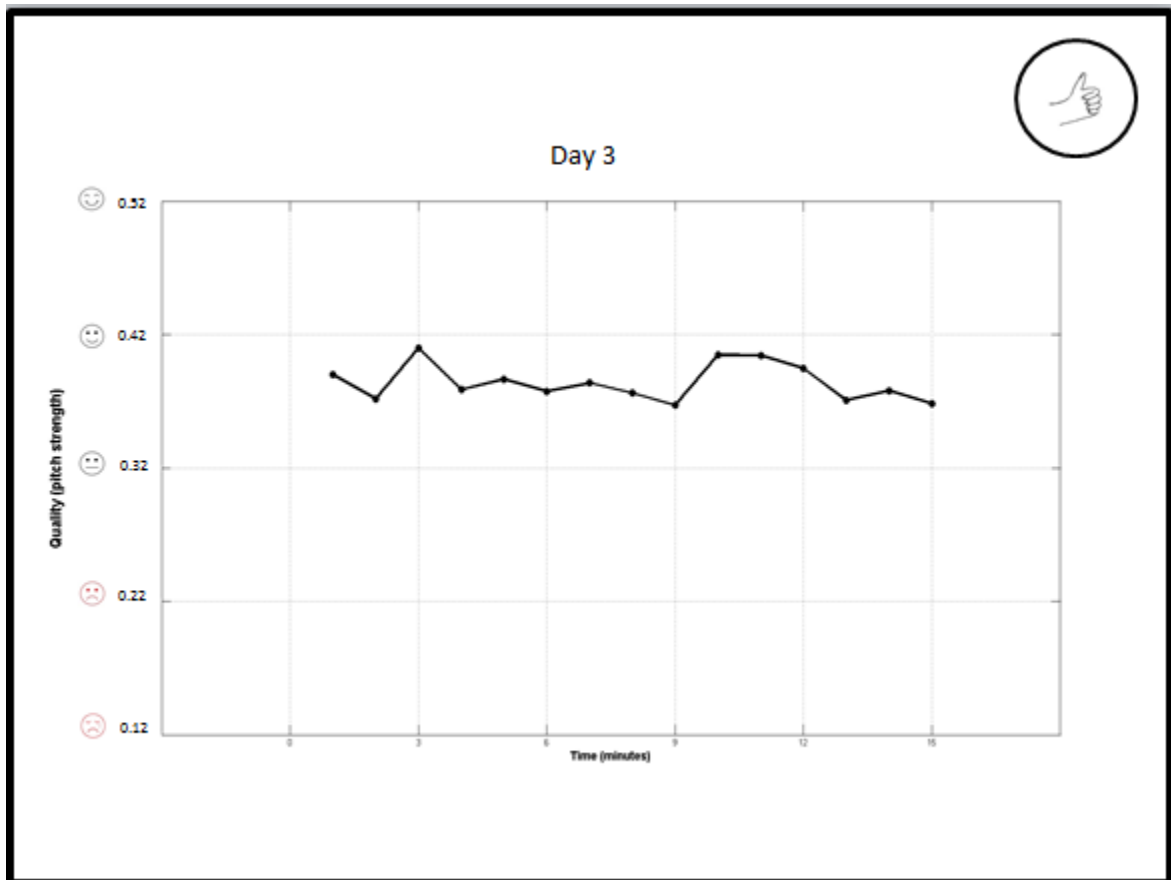
Figure 132: Third quality display. Amplified image of one day's quality. Participants had one of these for each day.
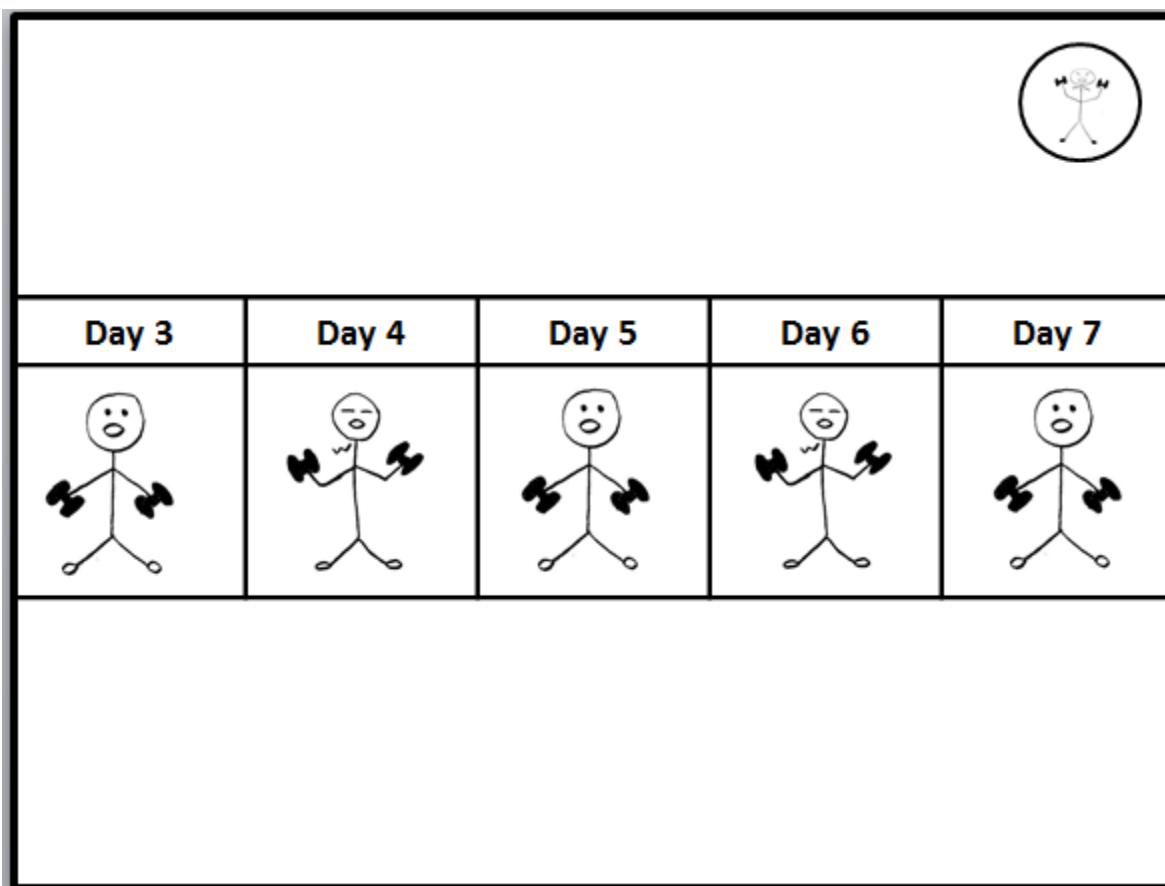
Figure 133: First strain display. The stick figures indicate the average value for the /ɑ/ vowels after the reading VLT (whichever value they are closest to). Note that the stick figure for days 3, 5, and 7 is the "baseline or better" stick figure.
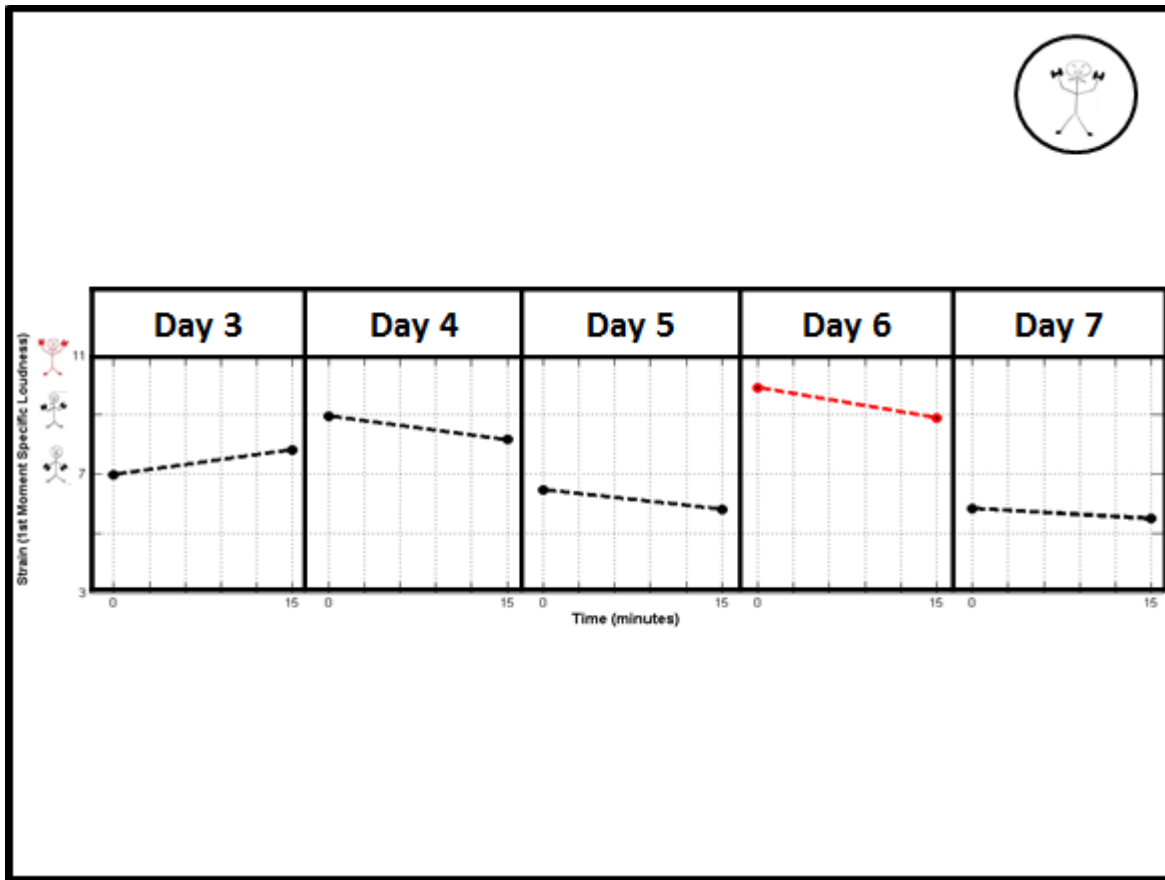
Figure 134: Second strain display. Note that if the value went above the second stick figure, it was considered to be in the "danger zone" and was colored red.

"Baseline" was defined as the average value for the three values before the VLT for the three baseline recordings. The range around the baseline was based on the range of first moment specific loudness values seen for voices ranging from no perceived strain to high strain (Kopf et al., 2013), where the value for a voice with low strain (10.9) was approximately 8 points lower than the voice with the highest strain (19.5). This range was chosen to ensure that all values for a given participant would fall within that range (e.g., even if fatigue leads to an increase in strain, it should still fall within this range).
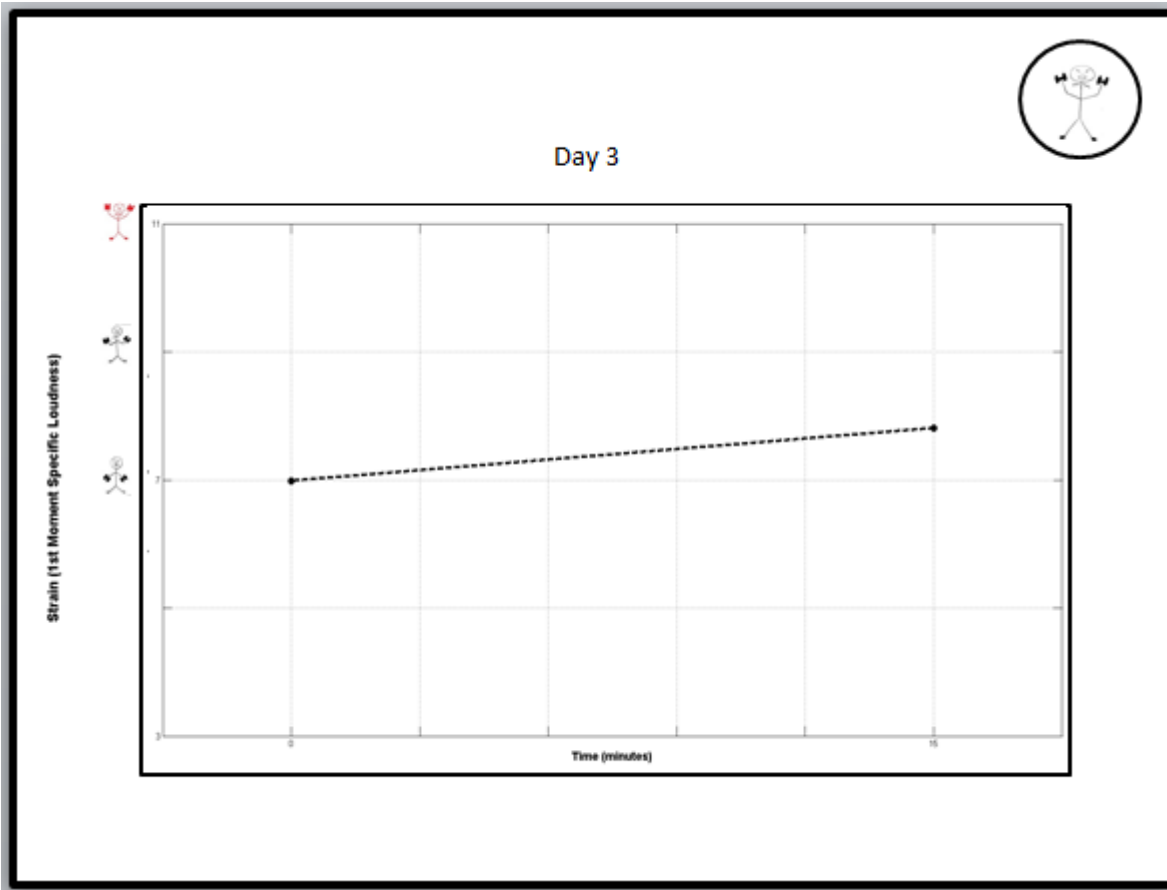
Figure 135: Third strain display. Amplified image of one day's strain. Just like for quality, participants had one of these for each day.

Appendix L: Supplemental Feedback Displays (Part 2.2)
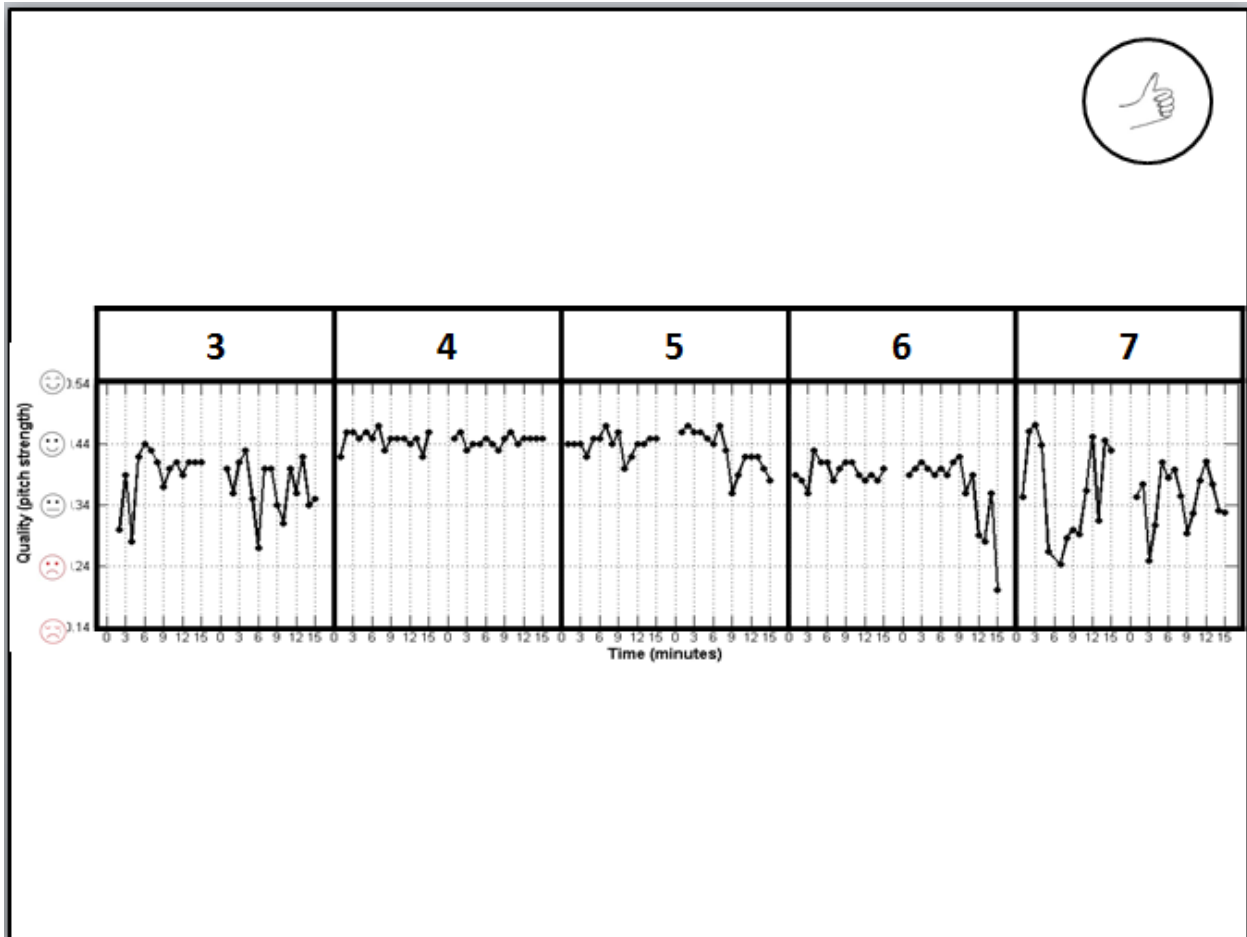
Figure 136: Example quality display. This display shows the difference between the quality display in Part 2.1, where instead of one 15-minute segment, two 15 minute segments are shown.

To keep the feedback consistent across participants, the first and last 15 minutes of the course were used (regardless of whether long periods of silence were present). Participants were asked to take that into consideration when looking at these graphs, which may explain some of the fluctuations present. The "zoomed in" version would be similar to the other third quality graph, where only one day is displayed.

Day 1	Day 2	Day 3	Day 4	Day 5
0.35 Minutes in Danger Zone	0.5 Minutes in Danger Zone	0.9 Minutes in Danger Zone	0.25 Minutes in Danger Zone	0 Minutes in Danger Zone

Figure 137: First loudness display. The "danger zone" differs by day. It represents time spent greater than 2 standard deviations of the mean above the average dB level for that day.
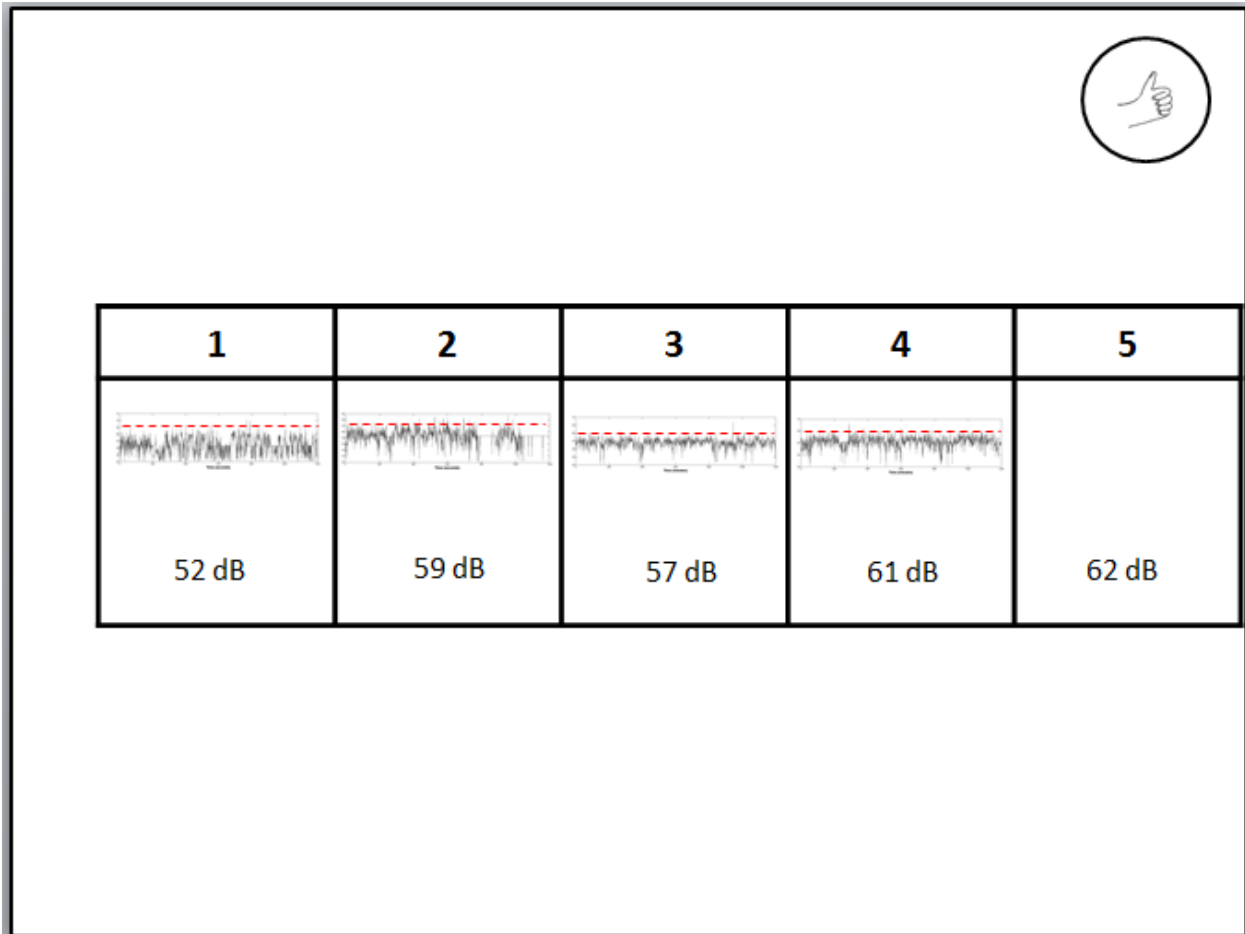
Figure 138: Second loudness display. The "danger zone" is indicated by the red dashed lines.

For display purposes, each data point in these graphs represents the average dB level for a three second time window. Note that for display purposes, silences were not marked by zero values, but by the average value for that day. This insured that the feedback image looked more similar to the final feedback image from Part 1.
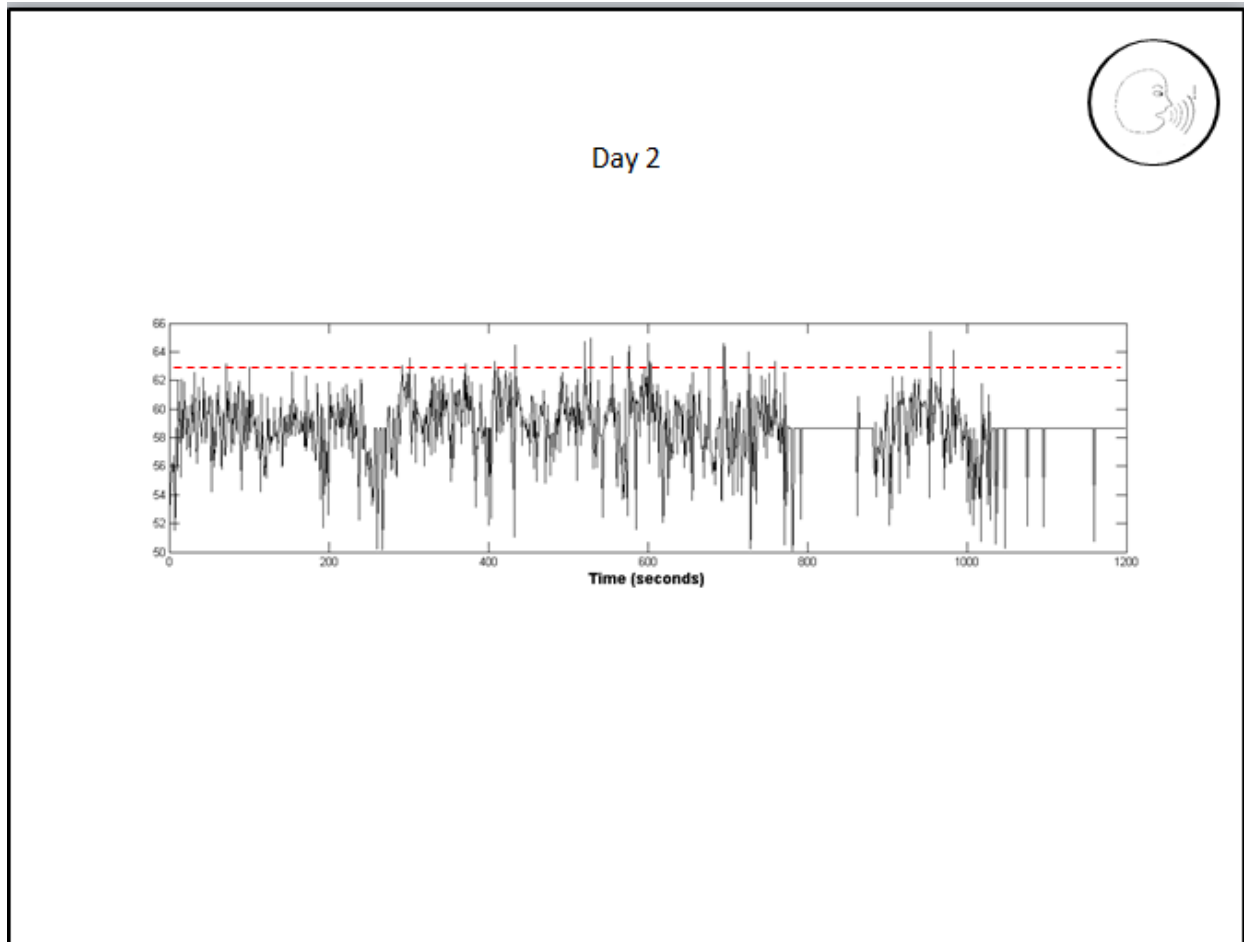
Figure 139: Third loudness display. Amplified image of one day's loudness pattern. Just like for quality, participants had one of these for each day.

Appendix M: Part 2.1 Readiness to Change
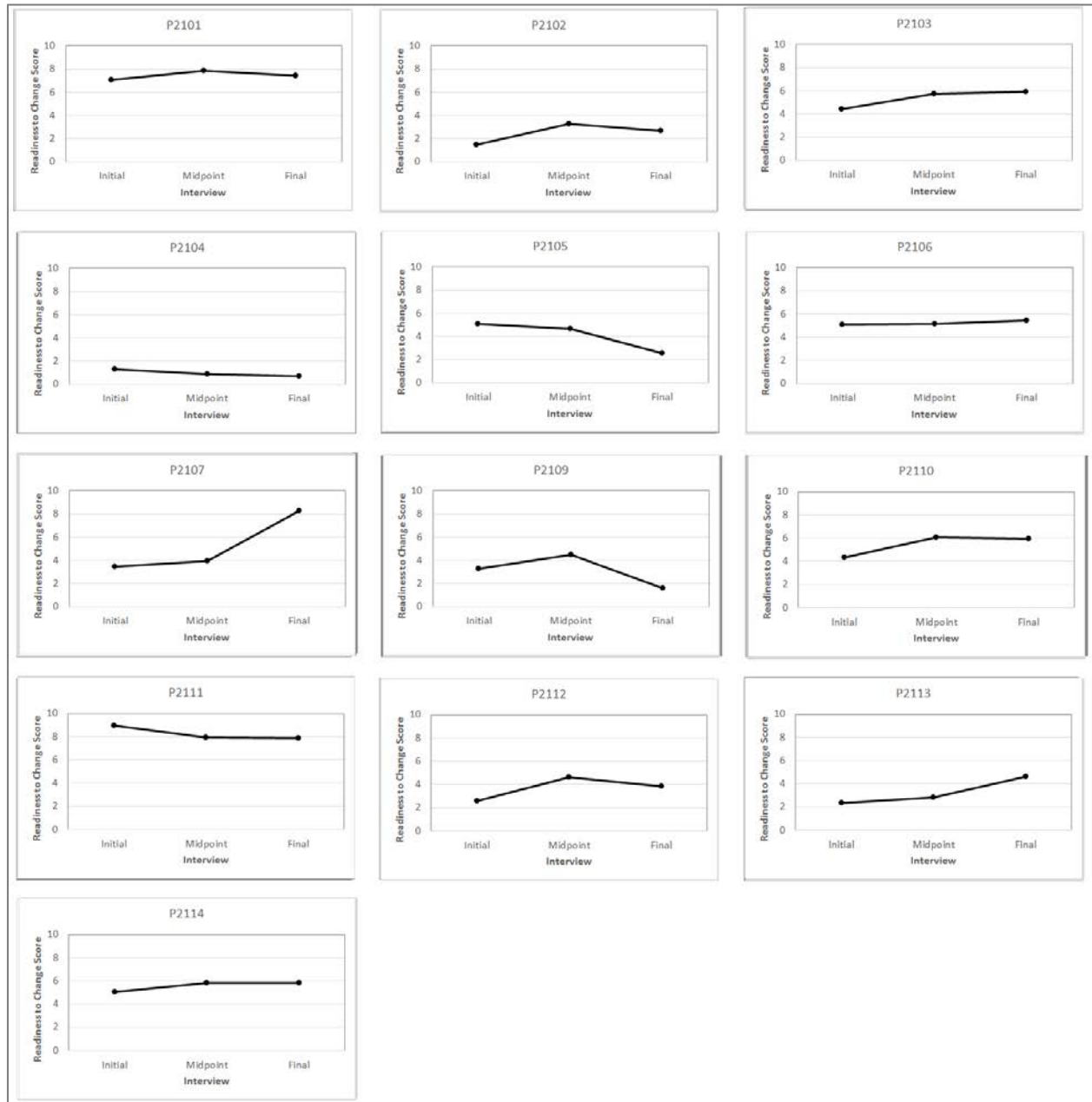
Figure 140: The change in readiness to change for each participant in Part 2.1 from the initial to the midpoint to the final interview.
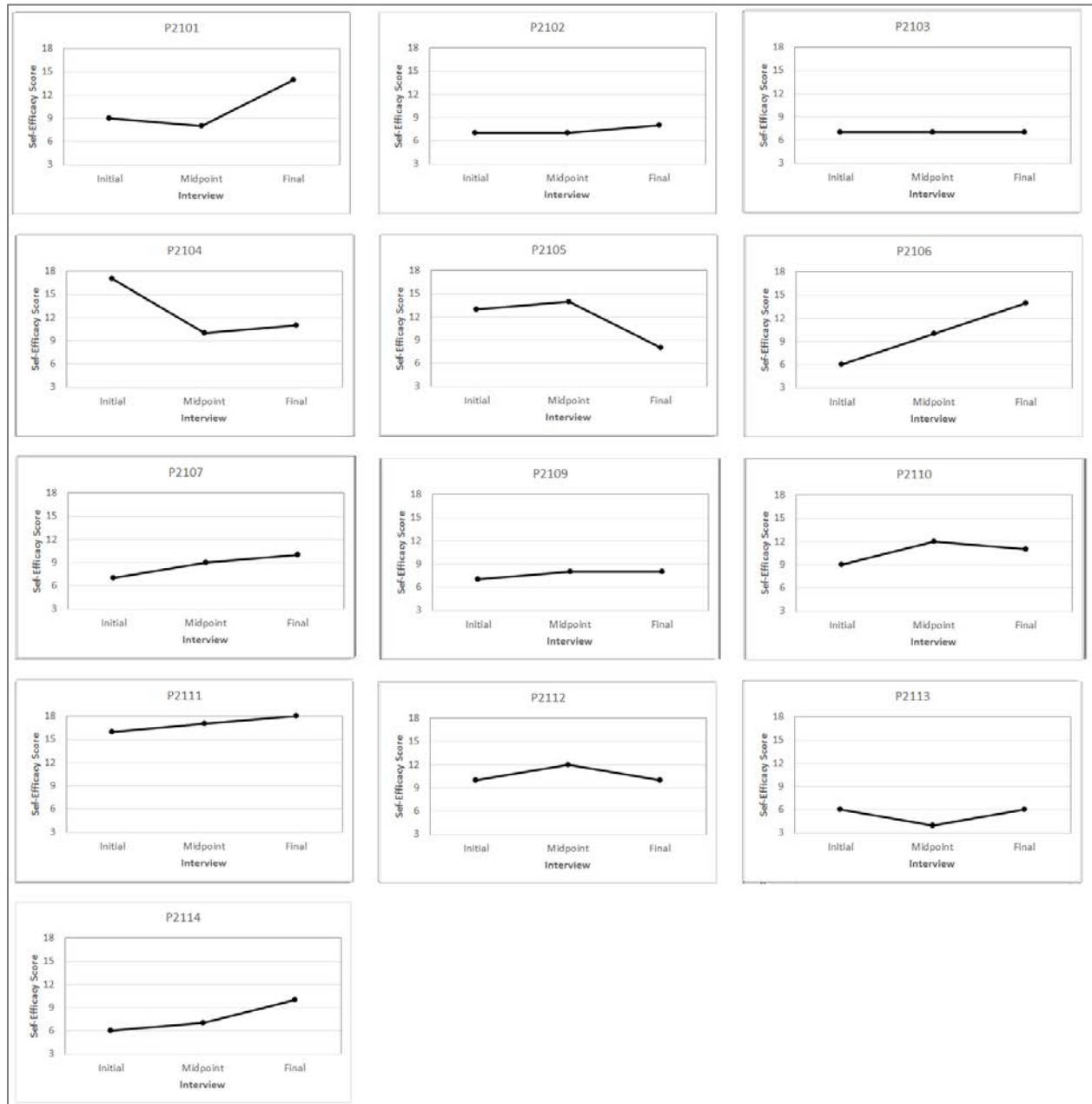
Figure 141: The change in self-efficacy for each participant in Part 2.1 from the initial to the midpoint to the final interview.

Appendix O: Part 2.1 Vocal Fatigue Index
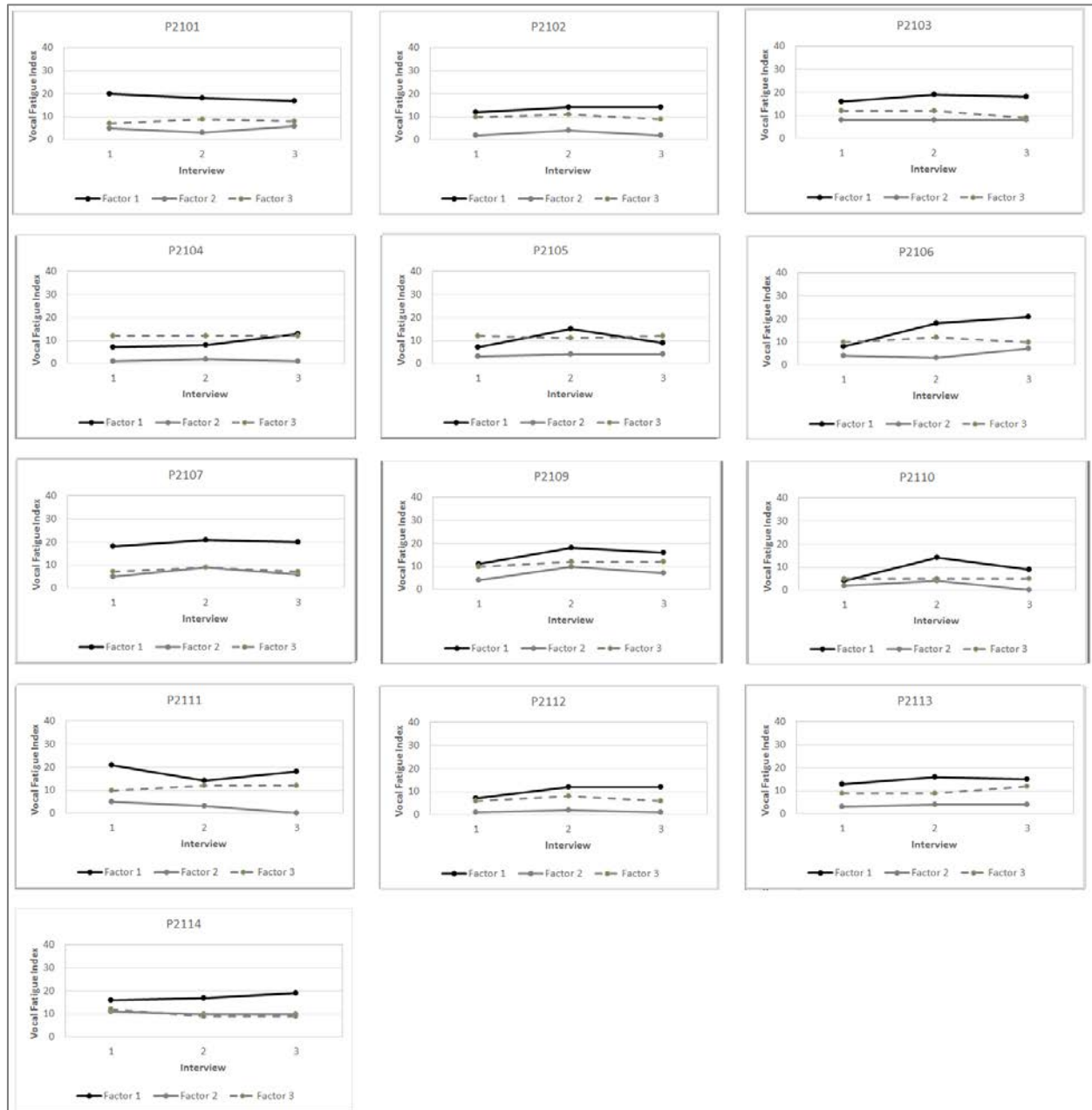
Figure 142: The change in vocal fatigue for each participant in Part 2.1 from the initial to the midpoint to the final interview.

Appendix P: Part 2.2 Readiness to Change
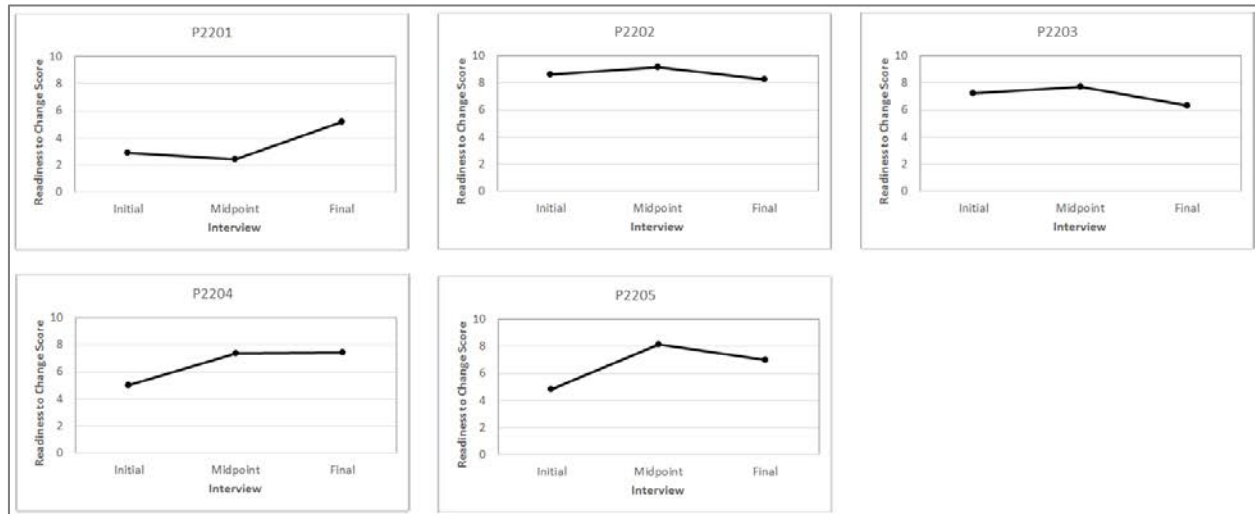
Figure 143: The change in readiness to change for each participant in Part 2.2 from the initial to the midpoint to the final interview.

Appendix Q: Part 2.2 Self-Efficacy
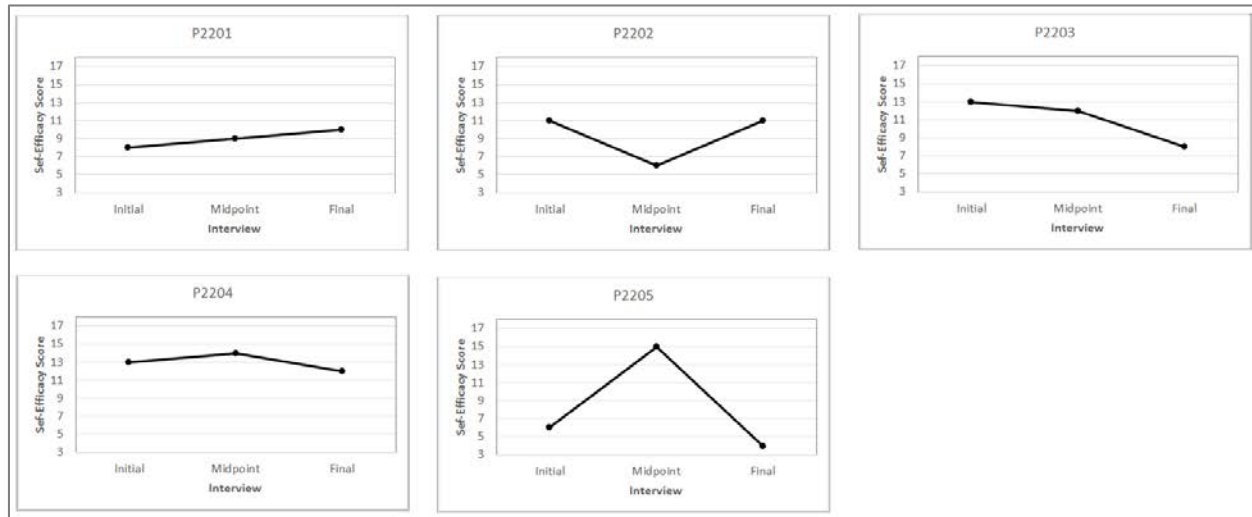
Figure 144: The change in self-efficacy for each participant in Part 2.2 from the initial to the midpoint to the final interview.

Appendix R: Part 2.2 Vocal Fatigue Index
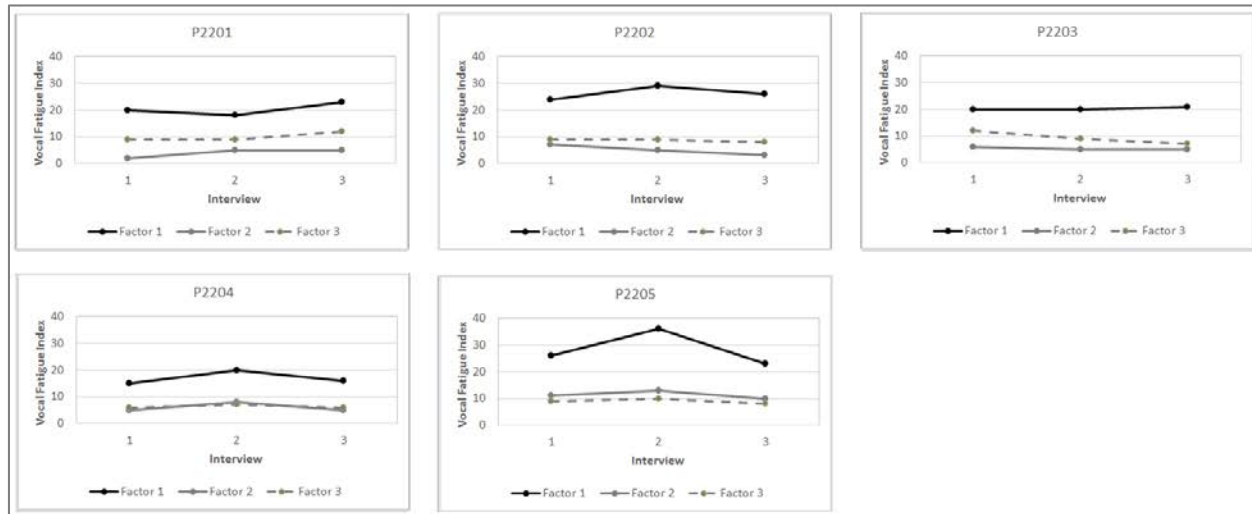
Figure 145: The change in vocal fatigue for each participant in Part 2.2 from the initial to the midpoint to the final interview.

Appendix S: Steps for Feedback Analysis (Part 2.1)

Step 1 Directions

1. Open GoldWave

2. Go to: File -> Batch Processing

3. Open the folder FullFiles

4. Highlight all files and drag into Batch Processing window

5. At the bottom of the Batch Processing window, under Presets, choose SplitAccData

6. In the Batch Processing window, click Begin

7. Once it finishes, click OK

8. The program has now put the correct channel from the recording into the SplitFiles folder

9. Open the SplitFiles folder and rename the recording, substitute "Acc" for "Roland" so that it looks like this: P2101_D1_Acc

10. Copy and paste the file into the Step 2 folder

11. Move the audio files from FullFiles & SplitFiles into the Complete folder

Step 2 Directions

1. Double click on ExtractdBF0 (MATLAB Code)- this will open MATLAB & MATLAB editor.

2. Right click on the wav file and choose rename. Copy the name of the file (but don't change it).

3. In the MATLAB editor window, replace fname- line 5 (paste the name of the wav file but keep .wav at the end).

4. In the MATLAB editor window, replace namexls- line 6 (paste the name of the wav file but keep .xlsx at the end).

5. Save the file (click the disk icon in the upper left).

6. Click run (big green arrow).

7. Drag and drop the audio file into Goldwave.

8. Go up to Effect- Volume- Change Volume and increase the volume by 20 dB (change 0.00 in upper right corner to 20). Click OK.

9. Now locate the 3 ah vowels toward the beginning of the recording.

10. Find the start and stop time for the 3rd ah vowel.

11. Close Goldwave and don't save changes.

12. Open the Excel file that you just created.

13. Find the start time for the 3rd ah vowel and note the corresponding SPL (sound pressure level).

14. In an empty cell next to it, type:

    = average (highlight SPL values for the duration of the 3rd ah vowel)

15. Now, in another empty cell, type:

    =(value from the Actual dB Excel sheet) – (highlight the cell where you did the averaging)

16. Go into the Step 1 Complete folder and copy both the full (Roland) and accelerometer (acc) recordings. Paste the copies into the Step 2 ChangeLevel folder.

17. Open Goldwave, drag and drop one of the files into the program.

18. Go up to Effect- Volume- Change Volume and increase the volume by the number of dB you got for 18. (Change 0.00 in upper right corner to that number). Click OK. If it was more than 20, add 20 and then add the rest by doing this again.

19. Save the file to: Y:\Graduate students\PhD students\LisaDissertation\Step2-AdjustdB\Complete

    Rename the file by adding "_cal" to the end (without quote marks).

20. Do the same for the other file.

Step 3 Directions

1. Use the 1cc_inc files that are in this folder.

2. Open and segment a "_Roland_cal" wav file in Elan (see Elan Instructions- Word file).

3. Close MATLAB if it's already open.

4. Double click CreateSegmentedWaves (MATLAB file) to open MATLAB.

5. Close the Editor window.

6. Type CreateSegmentedWaves (or copy paste it from here) into the MATLAB Command Window.

7. Follow the prompts- choose the ELAN file you just created, and **choose the corresponding "_acc_cal" wav file** to be segmented.

8. Once the wav files are created, rename all of them (take ".wav" off the end of each- it's redundant; change the one ending in "xxxxx" to the appropriate label).

9. Put all files that you used (wav & Elan) in the Complete folder under the correct participant and session.

ELAN Instructions

1. Open ELAN
2. Go to "File" → "New," select the .wave file you want to work on

3. You can listen to the wave form by clicking on it and pressing "play"
    a. After you've pressed "play" once, you can play and pause by pressing "space"
4. You can slow down the speed of the playback by moving the "rate" slider in the controls tab by moving left or right
5. "Options"→"Segmentation mode"
    At the top, choose "Two keystrokes per annotation"- this way you won't have to delete any.
    a. Click on waveform and press "enter" to the first segment
    b. Press "play" and hit "enter" when you want to end the segment
        i. You will be pressing "enter" to both start a new segment and end a segment
    c. Press "space" when you want to make the waveform stop playing all together
6. "Options"→"Transcription Mode" (Once you've segmented the entire passage)
    a. IT will bring each segment up (Tier)
    b. Select the tier name and click "apply"
    c. It will show all segments in that tier
    d. Name them using the following:
        Note: sometimes I ask someone more than 3 ah vowels if the first ones are short. If there are more than 3, go with the last 3 in the set.
        ia1 (first a vowel before reading), ia2, ia3, irp (initial rainbow passage), red (reading), ea1 (first a vowel after reading), ea2, ea3, erp (end rainbow passage)
7. Once all the segments and tiers are transcribed, you go to "File"→"Export As"→"Tab-delimited text…"
    a. Name it using the participant # and session # (e.g., P2102_D8)
    b. That will give you a .txd file with "begin time," "end time," "duration"
8. Go to Save and save using the same name from a.

Step 4 Directions

1. Copy all the /ɑ/ vowels from Step 3 into Midpoint500- Vowels.

2. Rename all ah vowels. Get rid of the .wav at the end. (It's on there 2x)

3. Double click the Midpoint500 MATLAB file to open it.

4. Click the green arrow to run it.

5. Once it is complete, go to the Vowels folder, rearrange the contents by Date Modified, highlight all the files ending in _mid, **cut** and paste allStimuli folder.

6. Double click on the sharpnessExperiment MATLAB file.

7. Scroll down to line 112 and change strainSharpnessMoments_... to

   strainSharpnessMoments_0121 (or whichever month and day you are running the

   analysis).

8. Save the script (upper left save icon).

9. Run the script (green arrow).

10. When finished, **cut** files from the Vowels folder and put in the Done folder. Also, **cut**

    files from allStimuli and put in the Done folder.

Step 5:

1. Run Auditory-SWIPE' (which calculates pitch strength) on the "red" file from Step 3.

2. Run MATLAB script to automatically calculate pause frequency and length of pauses

   >1 second based on output from SWIPE. Output will be a P21xx_Dx_pauses.xlsx

   file.

3. Open the P21xx_Dx_pauses.xlsx file with Excel.

4. Open the corresponding P21xx_PauseLocale.xlsx file.

5. [Note: these files should be open side by side to make your work easier!]

6. In the P21xx_Dx_pauses.xlsx file, in cell E1, paste the following:

   =COUNTIF(A:A,1).

7. [Note: this will tell you how many pauses occurred in the 15 minutes this person was

   reading]

8. In P21xx_Dx_pauses.xlsx, hold down the Ctrl button and push F (this will open the

   Find window).

9. In the text box, type the following: 1

10. Push the button Find Next.

11. On the line containing the 1 in Column A, there will be numbers in column B & C.

12. Copy both of these cells (Columns B & C), and paste them in the P21xx_PauseLocale.xlsx file. They will be pasted into columns A & B. If the number in column C is a 0, replace it with a 1 in the P21xx_PauseLocale.xlsx file.

13. Go back to the P21xx_Dx_pauses.xlsx file and double click on an empty cell next to the cells you just copied.

14. Click the Find Next button in the Find and Replace window. And continue for the rest of the lines containing 1 (the number in cell E1).

15. To double check your work, when you reach the end, check the number in E1 of P21xx_Dx_pauses.xlsx with I1 of P21xx_PauseLocale.xlsx. They should be the same number.

16. Once you have copied pasted all of the cells, you can close P21xx_Dx_pauses.xlsx.

17. Next, you will be adding the pauses together in time segments. First, you will add together all the pauses that occurred in minutes 0-2.999 (column A). To do this, type in cell F2: =SUM(and then highlight the corresponding pause lengths in column B. For example, if all the pauses between 0-2.999 minutes were in rows 2-5, your formula in F2 would look like this: =SUM(B2:B5).

18. Continue for each of the time segments.

Step 6:

1. Create feedback displays using MATLAB scripts and Power Point template.

Appendix T: Steps for Feedback Analysis (Part 2.2)

Step 1 Directions

1. Open GoldWave

2. Go to: File -> Batch Processing

3. Open the folder FullFiles

4. Highlight all files and drag into Batch Processing window

5. At the bottom of the Batch Processing window, under Presets, choose SplitAccData

6. In the Batch Processing window, click Begin

7. Once it finishes, click OK

8. The program has now put the correct channel from the recording into the SplitFiles folder

9. Open the SplitFiles folder and rename the recording, substitute "Acc" for "Roland" so that it looks like this: P2201_D1_Acc

10. Move the audio files from FullFiles & SplitFiles into the Complete folder

11. Open Goldwave, drag and drop one of the accelerometer files into the program.

12. Go up to Effect- Volume- Change Volume and increase the volume by 20 dB.

13. Save the file to: Y:\Graduate students\PhD students\LisaDissertation\Step2-AdjustdB\Complete

    a. Rename the file by adding "_inc" to the end (without quote marks).

14. Do the same for the other files.

15. If the entire session's recording is in one file (lecture + initial and final tasks), open the file in Goldwave, find the start and end for the lecture, and save as 3 separate files (1cc for initial tasks, 2cc for lecture, 3cc for final tasks).

16. For the 2cc file, divide into 1 hour segments in Goldwave, indicated by 2c1, 2c2, etc, and save to the Step 2 folder.

17. For the 2cc file, save the first 15 minutes of the file to the First 15 folder (name using this convention: P220x_Dx_2_first).

18. For the 2cc file, save the last 15 minutes of the file to the Last 15 folder (name using this convention: P220x_Dx_2__last).

Step 2 Directions

1. Note: this will be run for the 2cx files as well as the ia3 file created in Step 3 (for calibration purposes).

2. Double click on ExtractdBF0 (MATLAB Code) - this will open MATLAB & MATLAB editor.

3. Right click on the wav file and choose rename. Copy the name of the file (but don't change it).

4. In the MATLAB editor window, replace fname- line 5 (paste the name of the wav file but keep .wav at the end).

5. In the MATLAB editor window, replace namexls- line 6 (paste the name of the wav file but keep .xlsx at the end).

6. Save the file (click the disk icon in the upper left).

7. Click run (big green arrow).

8. In the ia3 Excel sheet, note the corresponding SPL (sound pressure level).

9. In an empty cell next to it, type:

    a. = average (highlight SPL values for the duration of the 3rd ah vowel)

10. Now, in another empty cell, type:

a. =(value from the Actual dB Excel sheet) – (highlight the cell where you did the averaging)

11. Use this number to "calibrate" the values in subsequent dB analysis, done in Excel.

12. Open the Excel sheet for the 2c1 file. Copy paste Columns D-Z from template for determining presence of voicing.

13. In column N, change the calibration value to the one from Step 2.11. Change SPL and F0 mins if needed (based on speaker and noise present in recording).

14. Copy columns D-Z and paste into the Excel sheets for 2c2, etc.

15. For dB level calculations, copy and combine values from Column V for all Excel files into a new worksheet.

16. Find average and standard deviation values. Determine cutoff for "Danger Zone" (2 standard deviations above mean).

17. Replace all "zeros" with the average value.

18. Run pause extraction MATLAB script for each of the 2cx files.

19. Use the "pauses" worksheet from each file. Sort by descending values in column 1 (so that all pauses are at the top.

20. Copy and combine values from Columns B and C for all Excel files into a new worksheet. They will be pasted into columns A & B.

21. Calculate count of Column A to determine the number of pauses >1 second.

22. Next, you will be adding the pauses together in time segments. Length of segments will vary depending on the length of the lecture. A total of 4-6 approximately equal segments will be displayed to the course instructor. First, you will add together all the pauses that occurred in a given segment. For example, if it is a 10 minute segment, you will add

together the values in Column B from time 0.0-9.999. To do this, type in cell F2: =SUM

(and then highlight the corresponding pause lengths in column B. For example, if all the

pauses between 0-9.999 minutes were in rows 2-50, your formula in F2 would look like

this: =SUM(B2:B50).

Step 3 Directions

1. Use the 1cc_inc and 3cc_inc files that are in this folder.

2. Close MATLAB if it's already open.

3. Double click CreateSegmentedWaves (MATLAB file) to open MATLAB.

4. Close the Editor window.

5. Type CreateSegmentedWaves (or copy paste it from here) into the MATLAB

   Command Window.

6. Follow the prompts- choose the ELAN file you just created, and **choose the**

   corresponding "_acc_inc" wav file to be segmented.

7. Once the wav files are created, rename all of them (take ".wav" off the end of each-

   it's redundant; change the one ending in "xxxxx" to the appropriate label).

8. Put all files that you used (wav & Elan) in the Complete folder under the correct

   participant and session.

ELAN Instructions

1. Open ELAN
2. Go to "File" → "New," select the .wave file you want to work on
3. You can listen to the wave form by clicking on it and pressing "play"
   a. After you've pressed "play" once, you can play and pause by pressing "space"
4. You can slow down the speed of the playback by moving the "rate" slider in the controls
   tab by moving left or right
5. "Options"→"Segmentation mode"
   a. At the top, choose "Two keystrokes per annotation"- this way you won't have to
      delete any.
   b. Click on waveform and press "enter" to the first segment

     c. Press "play" and hit "enter" when you want to end the segment
        i. You will be pressing "enter" to both start a new segment and end a segment
     d. Press "space" when you want to make the waveform stop playing all together

6. "Options"→"Transcription Mode" (Once you've segmented the entire passage)
     a. IT will bring each segment up (Tier)
     b. Select the tier name and click "apply"
     c. It will show all segments in that tier
     d. Name them using the following:
        i. Note: sometimes I ask someone more than 3 ah vowels if the first ones are short. If there are more than 3, go with the last 3 in the set.
        ii. ia1 (first a vowel before reading), ia2, ia3, irp (initial rainbow passage), red (reading), ea1 (first a vowel after reading), ea2, ea3, erp (end rainbow passage)

7. Once all the segments and tiers are transcribed, you go to "File"→"Export As"→"Tab-delimited text…"
     a. Name it using the participant # and session # (e.g., P2102_D8)
     b. That will give you a .txd file with "begin time," "end time," "duration"

8. Go to Save and save using the same name from a.

Step 4 Directions

1. Copy all the /ɑ/ vowels from Step 3 into Midpoint500- Vowels.

2. Rename all ah vowels. Get rid of the .wav at the end. (It's on there 2x)

3. Double click the Midpoint500 MATLAB file to open it.

4. Click the green arrow to run it.

5. Once it is complete, go to the Vowels folder, rearrange the contents by Date Modified, highlight all the files ending in _mid, **cut** and paste allStimuli folder.

6. Double click on the sharpnessExperiment MATLAB file.

7. Scroll down to line 112 and change strainSharpnessMoments_... to strainSharpnessMoments_0121 (or whichever month and day you are running the analysis).

8. Save the script (upper left save icon).

9. Run the script (green arrow).

10. When finished, **cut** files from the Vowels folder and put in the Done folder. Also, **cut** files from allStimuli and put in the Done folder.

Step 5:

1. Run Auditory-SWIPE' (to calculate pitch strength) on the _first and __last files from Step 1.

Step 6:

1. Create feedback displays using MATLAB scripts and Power Point template.
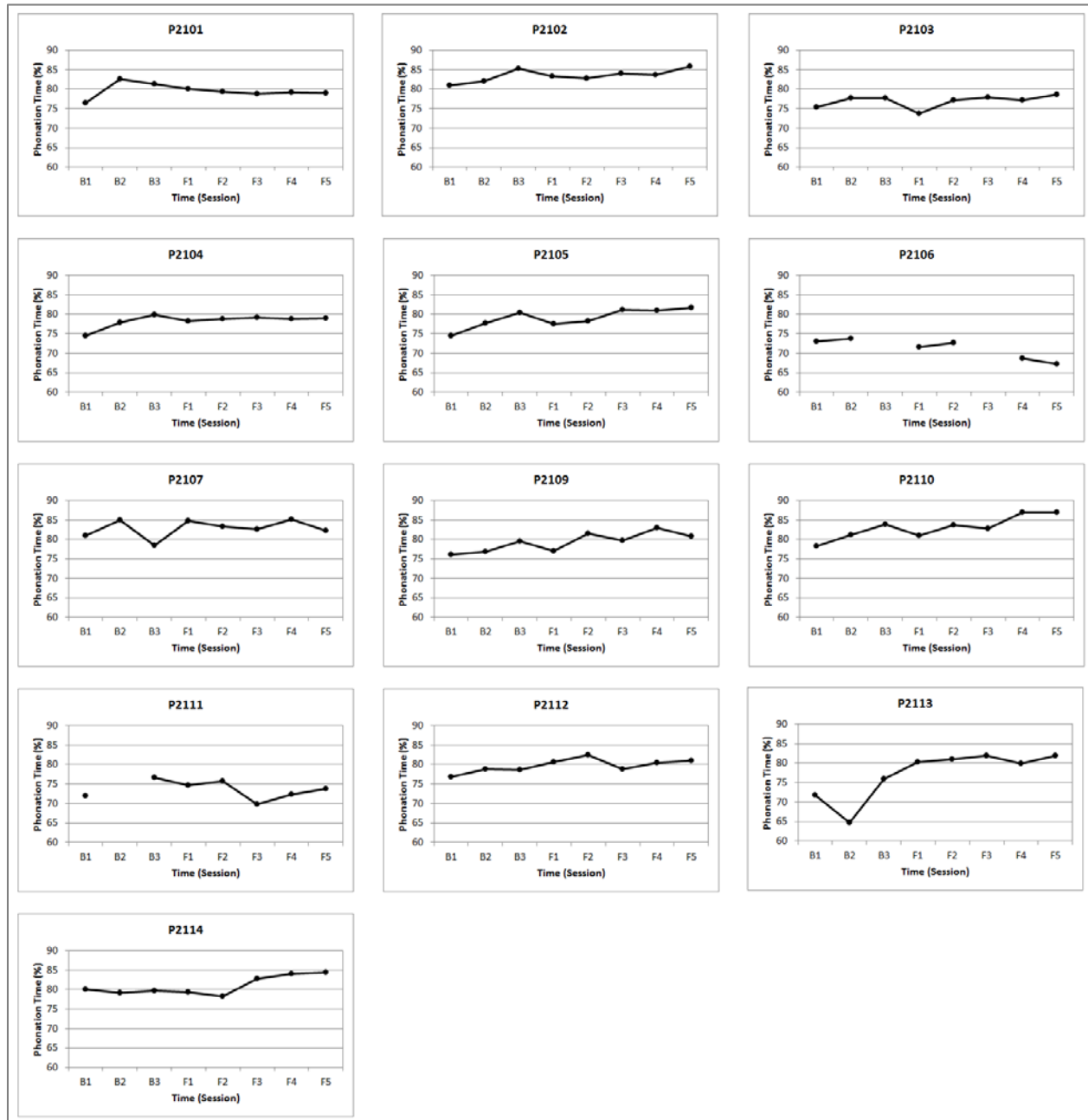
Figure 146: The change in average phonation time for each participant in Part 2.1 across the eight recording sessions. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix V: Part 2.1 Vocal Intensity
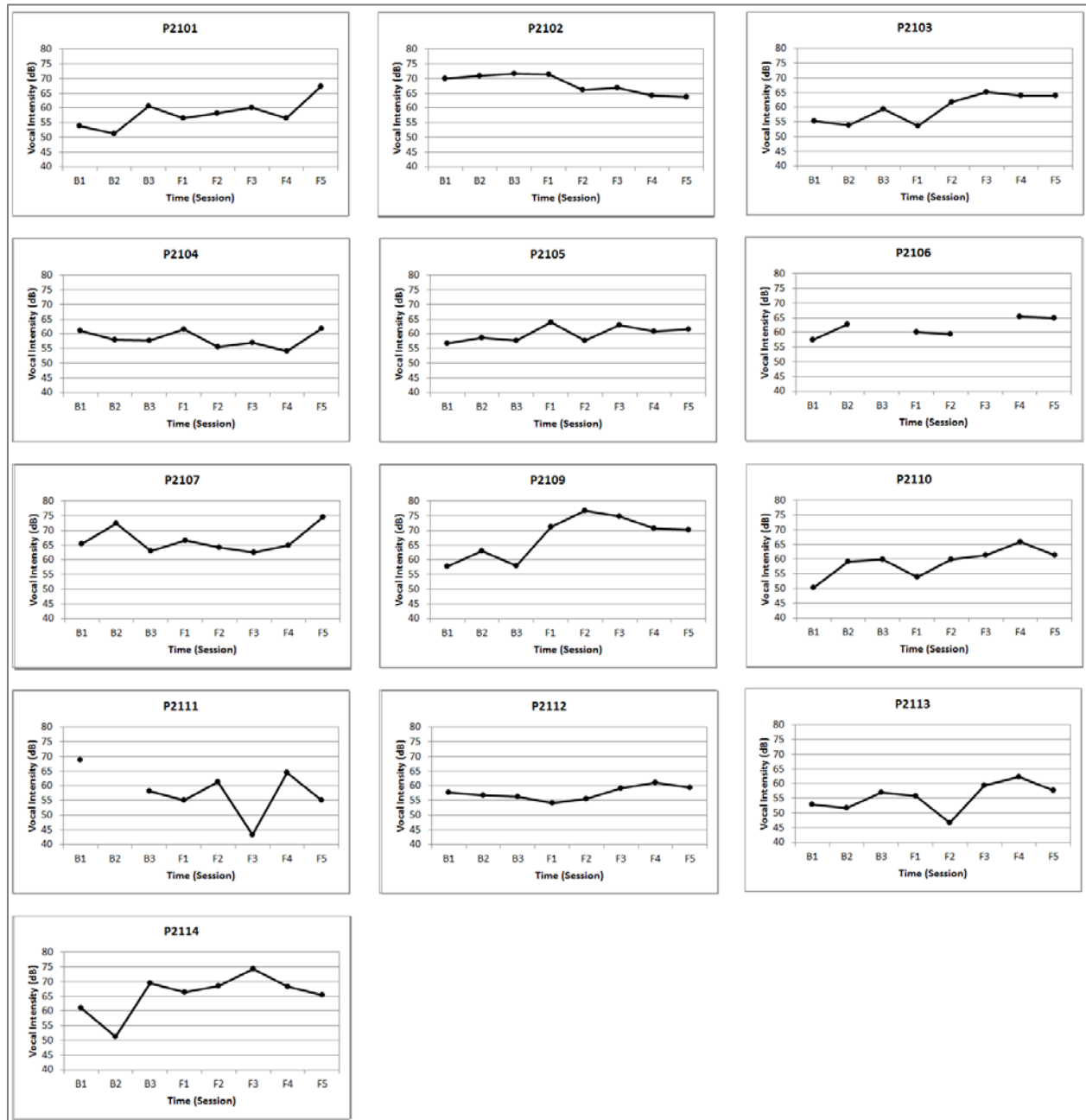
Figure 147: The change in average vocal intensity for each participant in Part 2.1 across the eight recording sessions. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix W: Part 2.1 Pitch
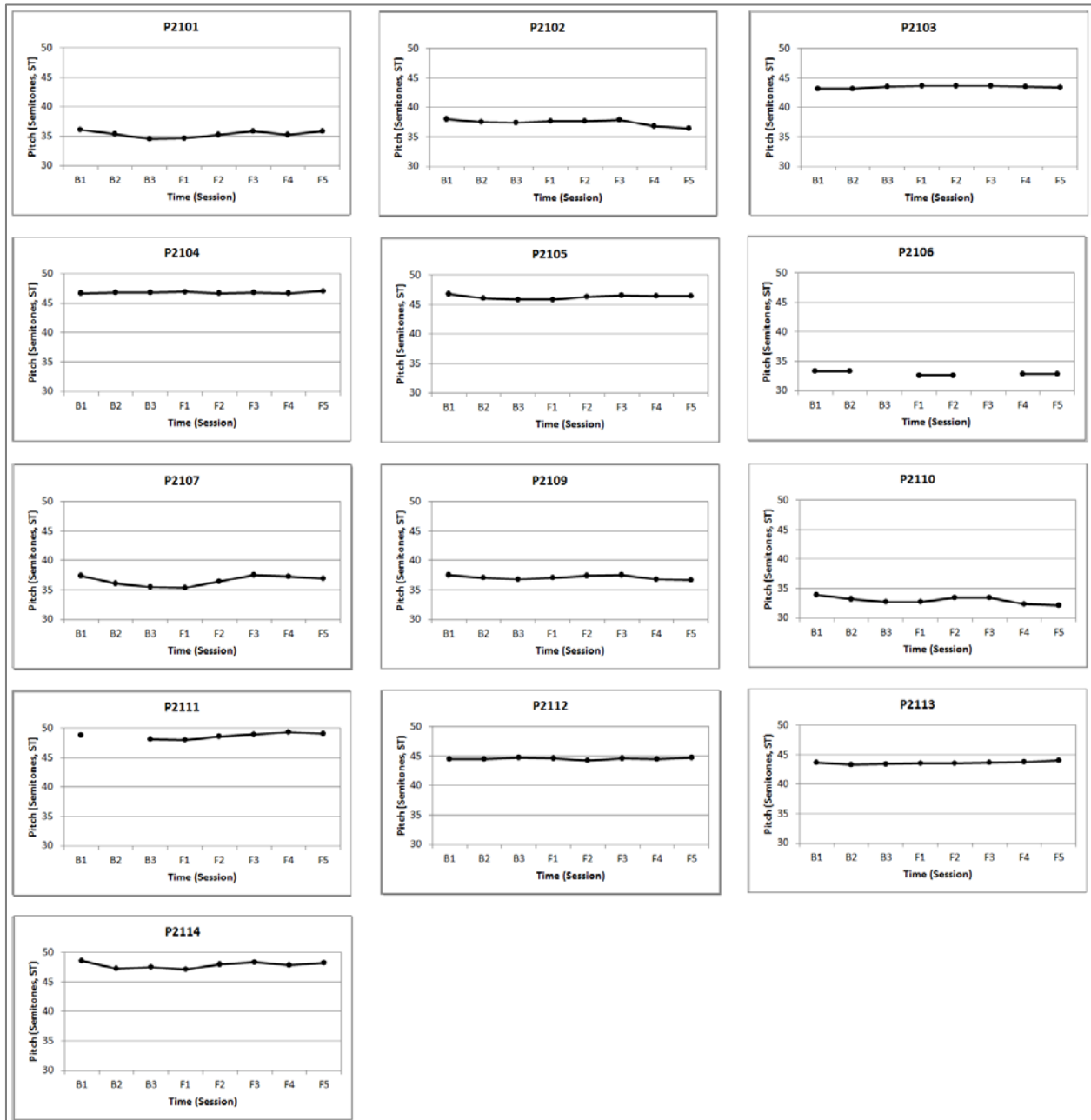
Figure 148: The change in average pitch for each participant in Part 2.1 across the eight recording sessions. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix X: Part 2.2 Phonation Time
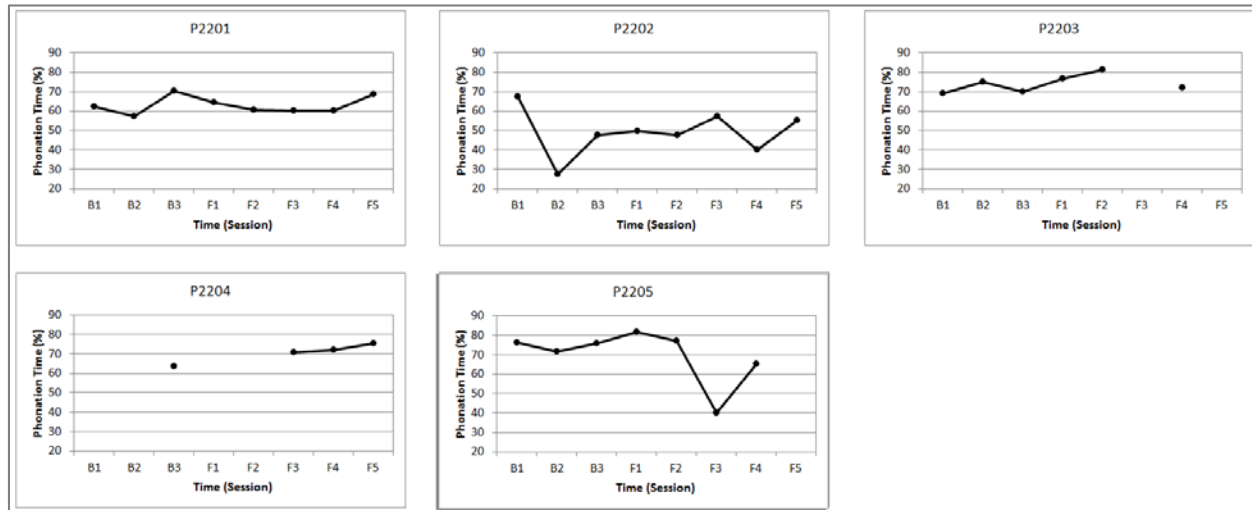
Figure 149: The change in average phonation time for each participant in Part 2.2 across the eight recording sessions. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix Y: Part 2.2 Vocal Intensity
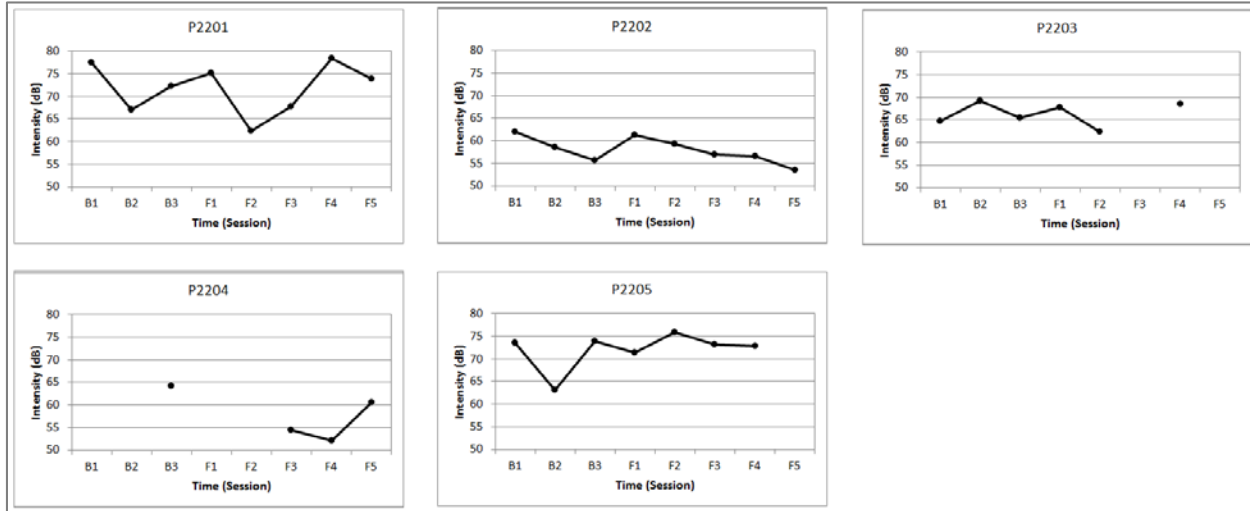
Figure 150: The change in average vocal intensity for each participant in Part 2.2 across the eight recording sessions. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix Z: Part 2.2 Pitch
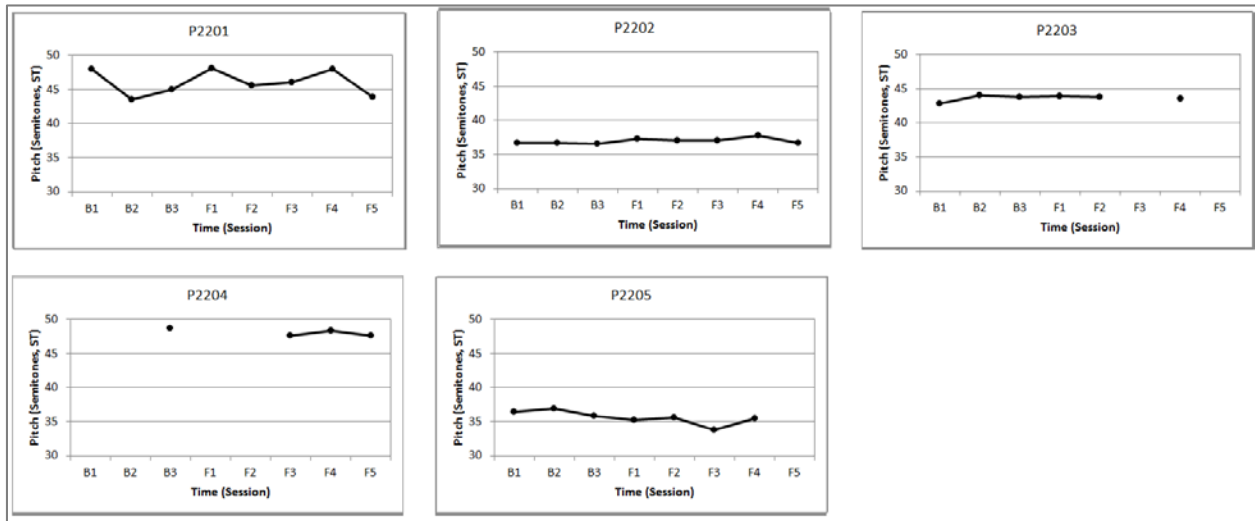
Figure 151: The change in average pitch for each participant in Part 2.2 across the eight recording sessions. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix AA: Average Pitch Strength (Part 2.1)
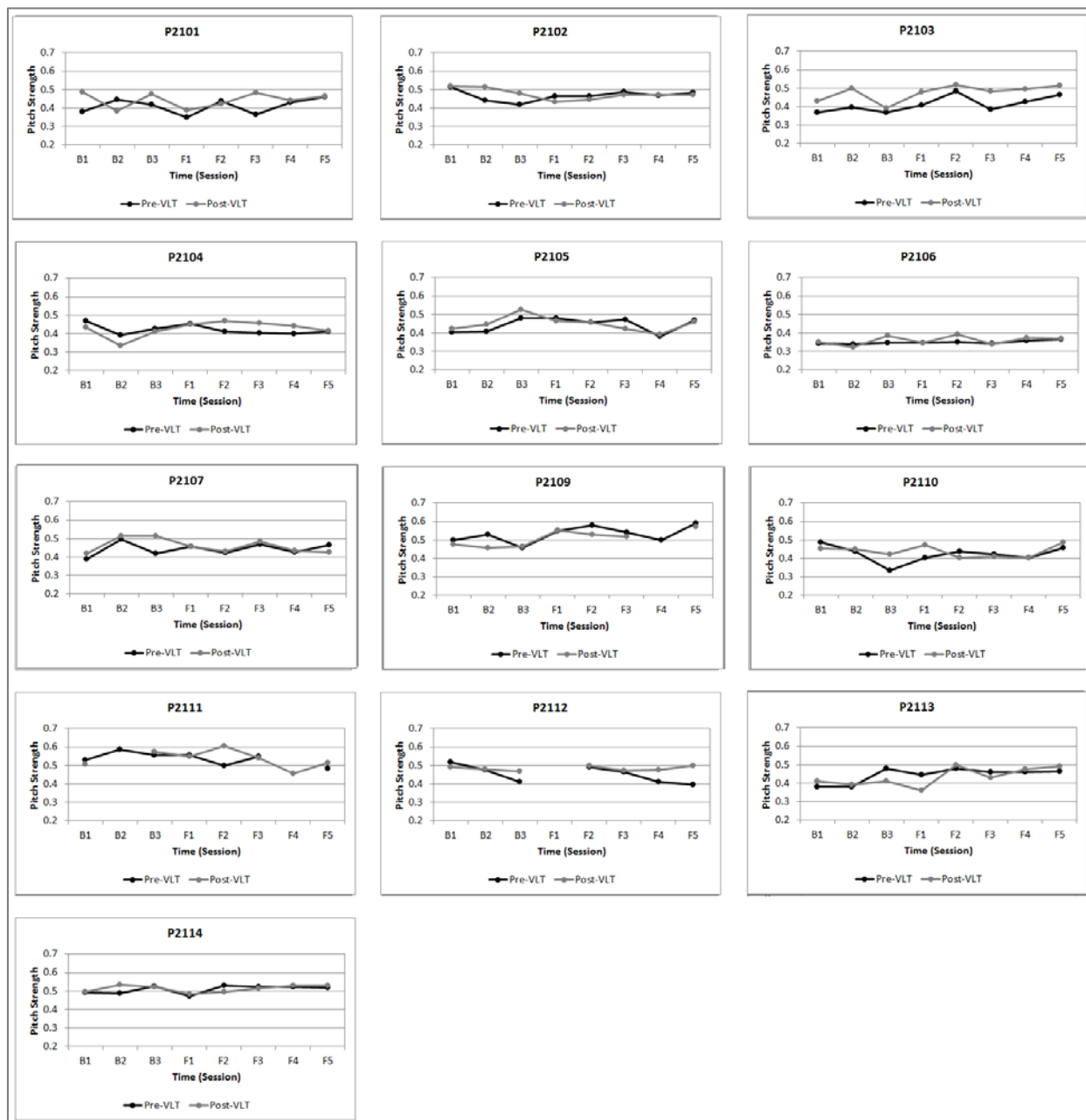
Figure 152: The average pitch strength for each participant in Part 2.1 before and after each vocal loading task. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix AB: Average First Moment Specific Loudness (Part 2.1)
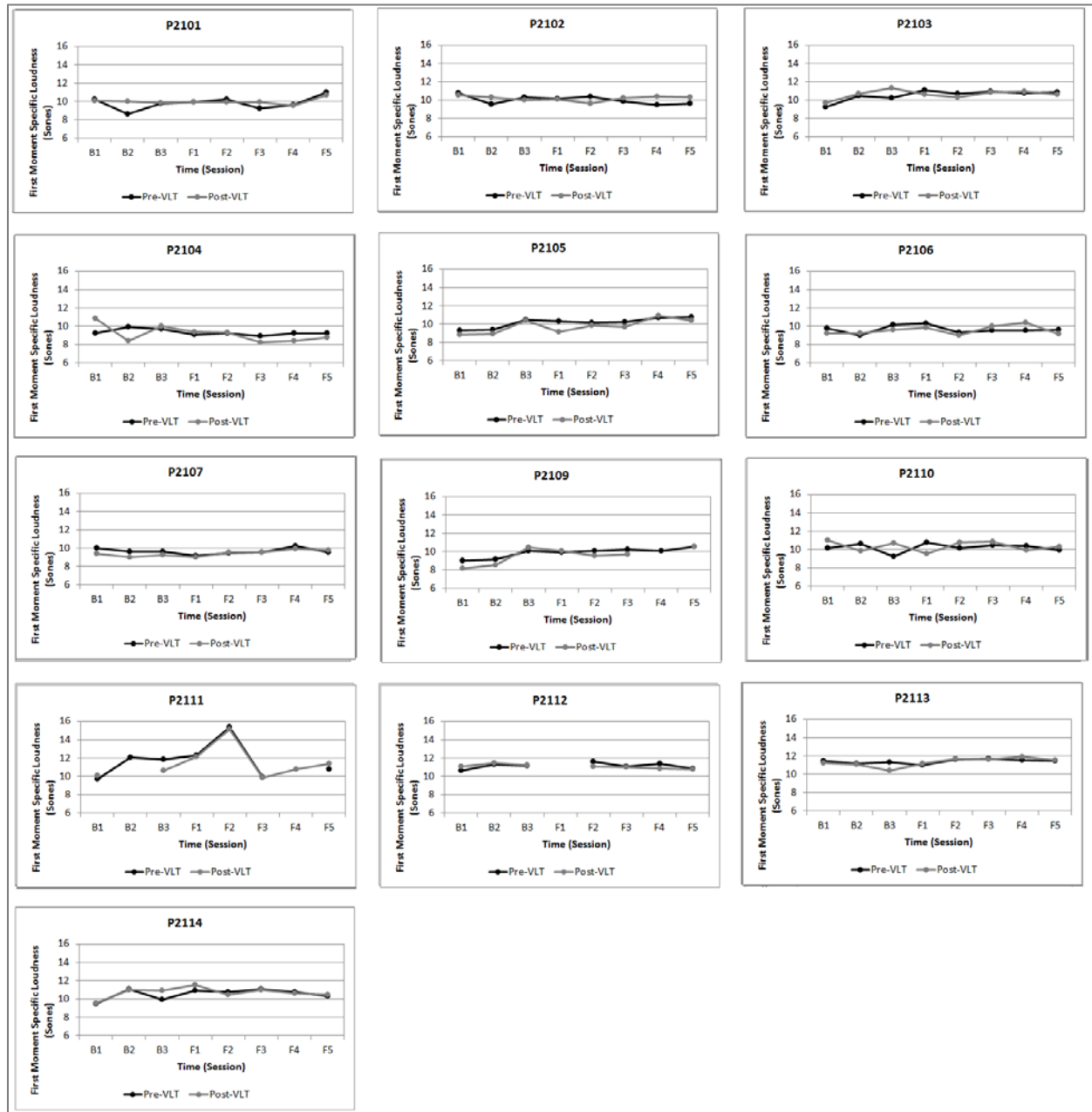
Figure 153: The average first moment specific loudness for each participant in Part 2.1 before and after each vocal loading task. B# indicates the baseline recording number, and F# indicates the feedback recording number.

Appendix AC: Average Pitch Strength (Part 2.2)
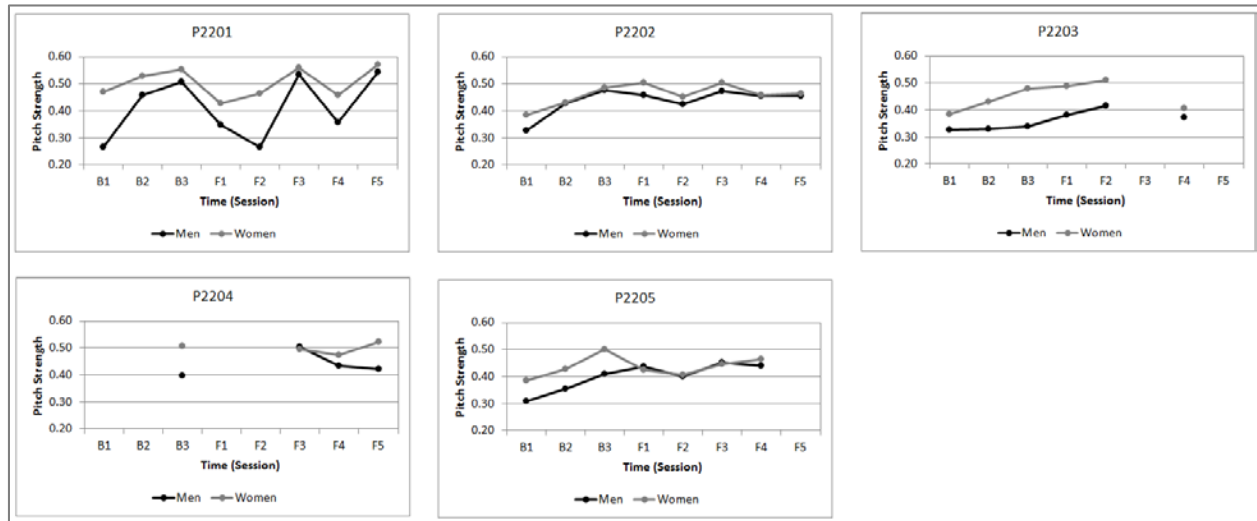
Figure 154: The average pitch strength for each participant in Part 2.2 before and after each vocal loading task. B# indicates the baseline recording number, and F# indicates the feedback recording number.
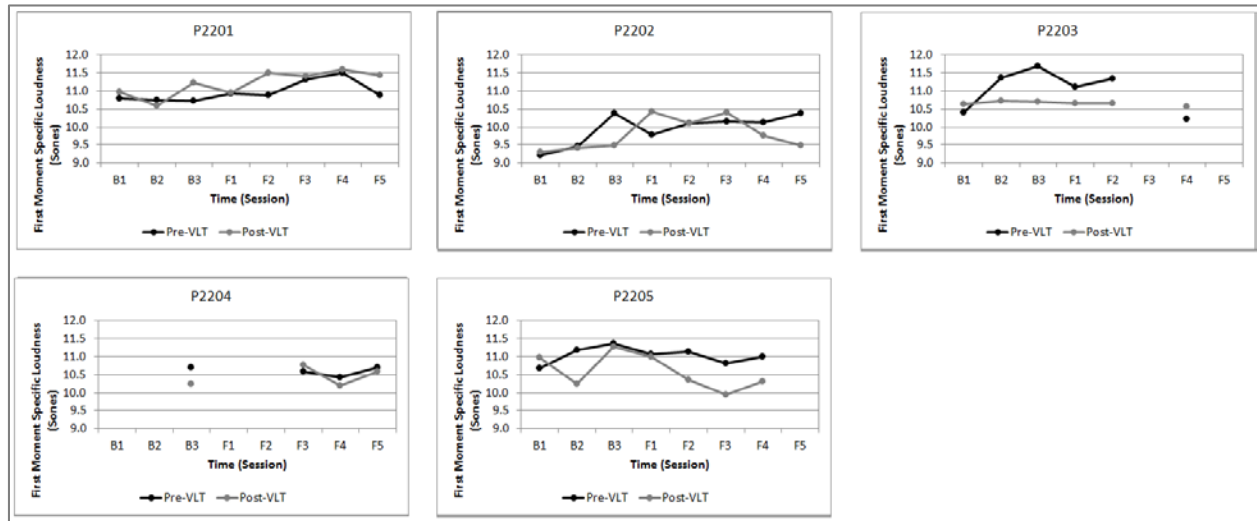
Figure 155: The average first moment specific loudness for each participant in Part 2.2 before and after each vocal loading task. B# indicates the baseline recording number, and F# indicates the feedback recording number.

BIBLIOGRAPHY

BIBLIOGRAPHY

Achey, M. A., He, M. Z., & Akst, L. M. (2016). Vocal Hygiene Habits and Vocal Handicap Among Conservatory Students of Classical Singing. *Journal of Voice*, *30*(2), 192–197. http://doi.org/10.1016/j.jvoice.2015.02.003

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18. http://doi.org/10.1037/a0021017

Astolfi, A., Carullo, A., Pavese, L., & Puglisi, G. E. (2015). Duration of voicing and silence periods of continuous speech in different acoustic environments. *The Journal of the Acoustical Society of America*, *137*(2), 565–579.

Astolfi, A., Carullo, A., Vallan, A., & Pavese, L. (2013). Influence of classroom acoustics on the vocal behavior of teachers. In *Proceedings of Meetings on Acoustics* (Vol. 19, p. 40123). Acoustical Society of America. Retrieved from http://scitation.aip.org/content/asa/journal/poma/19/1/10.1121/1.4800427

Bailey, G. "Skip." (1993). Iterative methodology and designer training in human-computer interface design. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (pp. 198–205). New York, NY, USA: ACM.

Baken, R. J. (1992). Electroglottography. *Journal of Voice*, *6*(2), 98–110. http://doi.org/10.1016/S0892-1997(05)80123-7

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191.

Bergan, C. C., Titze, I. R., & Story, B. (2004). The perception of two vocal qualities in a synthesized vocal utterance: ring and pressed voice. *Journal of Voice*, *18*(3), 305–317. http://doi.org/10.1016/j.jvoice.2003.09.004

Bernstein, L., & Yuhas, C. M. (2005). Prototyping. In *Trustworthy Systems through Quantitative Software Engineering* (pp. 107–136). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com.proxy1.cl.msu.edu/doi/10.1002/0471750336.ch4/summary

Beyer, H., & Holtzblatt, K. (1998). *Contextual Design : Defining Customer-Centered Systems*. San Francisco, Calif: Morgan Kaufmann. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=472251&scope=site

Bhuta, T., Patrick, L., & Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, *18*(3), 299–304. http://doi.org/10.1016/j.jvoice.2003.12.004

Bovo, R., Galceran, M., Petruccelli, J., & Hatzopoulos, S. (2007). Vocal problems among teachers: Evaluation of a preventive voice program. *Journal of Voice*, *12*(6), 705–722.

Boyle, R. G., O'Connor, P. J., Pronk, N. P., & Tan, A. (1998). Stages of change for physical activity, diet, and smoking among HMO members with chronic conditions. *American Journal of Health Promotion: AJHP*, *12*(3), 170–175.

Brant, L. J., & Fozard, J. L. (1990). Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging. *The Journal of the Acoustical Society of America*, *88*(2), 813–820. http://doi.org/10.1121/1.399731

Buley, L. (2013). *The User Experience Team of One*. Brooklyn, NY: Rosenfeld Media.

Camacho, A. (2007). *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. University of Florida. Retrieved from http://www.kerwa.ucr.ac.cr:8080/handle/10669/536

Cardinal, B. J. (1995). The stages of exercise scale and stages of exercise behavior in female adults. *The Journal of Sports Medicine and Physical Fitness*, *35*(2), 87–92.

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. http://doi.org/10.1016/j.otohns.2006.06.1268

Carullo, A., Vallan, A., & Astolfi, A. (2013). Design Issues for a Portable Vocal Analyzer. *IEEE Transactions on Instrumentation and Measurement*, *62*(5), 1084–1093. http://doi.org/10.1109/TIM.2012.2236724

Chan, R. W. (1994). Does the voice improve with vocal hygiene education? A study of some instrumental voice measures in a group of kindergarten teachers. *Journal of Voice*, *8*(3), 279–291.

Child, D. R., & Johnson, T. S. (1991). Preventable and nonpreventable causes of voice disorders. In *Seminars in Speech and Language* (Vol. 12, pp. 1–13). \copyright 1991 by Thieme Medical Publishers, Inc. Retrieved from https://www.thieme-connect.com/products/ejournals/pdf/10.1055/s-2008-1064206.pdf

Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data (pp. 1143–1152). ACM Press. http://doi.org/10.1145/2556288.2557372

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

Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., … others. (2008). Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1797–1806). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1357335

Corbin, J. M., & Strauss, A. L. (2008). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd ed.). Los Angeles, CA: Sage Publications, Inc.

Deliyski, D. D., & Hillman, R. E. (2010). State of the Art Laryngeal Imaging: Research and Clinical Implications. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *18*(3), 147–152. http://doi.org/10.1097/MOO.0b013e3283395dd4

DiClemente, C. C., Prochaska, J. O., Fairhurst, S. K., Velicer, W. F., Velasquez, M. M., & Rossi, J. S. (1991). The process of smoking cessation: An analysis of precontemplation, contemplation, and preparation stages of change. *Journal of Consulting and Clinical Psychology*, *59*(2), 295–304. http://doi.org/10.1037/0022-006X.59.2.295

DiClemente, C. C., Schlundt, D., & Gemmell, L. (2004). Readiness and stages of change in addiction treatment. *The American Journal on Addictions / American Academy of Psychiatrists in Alcoholism and Addictions*, *13*(2), 103–119. http://doi.org/10.1080/10550490490435777

Dijkstra, A., De Vries, H., & Bakker, M. (1996). Pros and cons of quitting, self-efficacy, and the stages of change in smoking cessation. *Journal of Consulting and Clinical Psychology*, *64*(4), 758–763.

Duffy, O. M., & Hazlett, D. E. (2004). The impact of preventive voice care programs for training teachers: A longitudinal study. *Journal of Voice*, *18*(1), 63–70. http://doi.org/10.1016/S0892-1997(03)00088-2

Eddins, D. A., & Shrivastav, R. (2013). Psychometric properties associated with perceived vocal roughness using a matching task. *The Journal of the Acoustical Society of America*, *134*(4), EL294-300. http://doi.org/10.1121/1.4819183

Epstein, D. A., Borning, A., & Fogarty, J. (2013). Fine-grained sharing of sensed physical activity: a value sensitive approach (p. 489). ACM Press. http://doi.org/10.1145/2493432.2493433

Fairbanks, G. (1960). *Voice and Articulation Drillbook* (2nd ed.). New York, NY, USA: Harper.

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and Models*. Springer.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. San Francisco, CA: Morgan Kaufmann Publishers.

Folkins, J. W., Brackenbury, T., Krause, M., & Haviland, A. (2015). Enhancing the Therapy Experience Using Principles of Video Game Design. *American Journal of Speech-Language Pathology*, 1. http://doi.org/10.1044/2015_AJSLP-14-0059

Fritz, T., Huang, E. M., Murphy, G. C., & Zimmermann, T. (2014). Persuasive Technology in the Real World: A Study of Long-term Use of Activity Sensing Devices for Fitness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 487–496). New York, NY, USA: ACM. http://doi.org/10.1145/2556288.2557383

Froeschels, E. (1952). Chewing method as therapy: A discussion with some philosophical conclusions. *Archives of Otolaryngology–Head & Neck Surgery*, *56*(4), 427–434.

Gartner-Schmidt, J. L., Roth, D. F., Zullo, T. G., & Rosen, C. A. (2013). Quantifying Component Parts of Indirect and Direct Voice Therapy Related to Different Voice Disorders. *Journal of Voice*, *27*(2), 210–216. http://doi.org/10.1016/j.jvoice.2012.11.007

Gee, J. P. (2008). *Good Video Games + Good Learning: Collected Essays on Video Games, Learning and Literacy*. New York, NY, USA: Peter Lang.

Ghassemi, M., Van Stan, J. H., Mehta, D. D., Zañartu, M., Cheyne, H. A., Hillman, R. E., & Guttag, J. V. (2014). Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules. *IEEE Transactions on Bio-Medical Engineering*, *61*(6), 1668–1675. http://doi.org/10.1109/TBME.2013.2297372

Goldman, N., & Narayanaswamy, K. (1992). Software Evolution Through Iterative Prototyping. In *Proceedings of the 14th International Conference on Software Engineering* (pp. 158–172). New York, NY, USA: ACM. http://doi.org/10.1145/143062.143109

Griffin, M. (2004). Minimum health and safety requirements for workers exposed to hand-transmitted vibration and whole-body vibration in the European Union; a review. *Occupational and Environmental Medicine*, *61*(5), 387–397. http://doi.org/10.1136/oem.2002.006304

Grillo, E. U., & Verdolini, K. (2008). Evidence for Distinguishing Pressed, Normal, Resonant, and Breathy Voice Qualities by Laryngeal Resistance and Vocal Efficiency in Vocally Trained Subjects. *Journal of Voice*, *22*(5), 546–552. http://doi.org/10.1016/j.jvoice.2006.12.008

Hansen, A. L. (1997). Reflections on I/Design: User interface design at a startup. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 487–493). New York, NY, USA: ACM.

Harrison, D., Berthouze, N., Marshall, P., & Bird, J. (2014). Tracking physical activity: problems related to running longitudinal studies with commercial devices (pp. 699–702). ACM Press. http://doi.org/10.1145/2638728.2641320

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. http://doi.org/10.1016/j.wocn.2010.08.002

Heman-Ackah, Y. D., Michael, D. D., Baroody, M. M., Ostrowski, R., Hillenbrand, J., Heuer, R. J., … Sataloff, R. T. (2003). Cepstral Peak Prominence: A More Reliable Measure of Dysphonia. *Annals of Otology, Rhinology & Laryngology*, *112*(4), 324–333. http://doi.org/10.1177/000348940311200406

Hillman, R. E., Heaton, J. T., Masaki, A., Zeitels, S. M., & Cheyne, H. A. (2006). Ambulatory monitoring of disordered voices. *Annals of Otology, Rhinology & Laryngology*, *115*(11), 795–801.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1989). Objective assessment of vocal hyperfunctionan experimental framework and initial results. *Journal of Speech, Language, and Hearing Research*, *32*(2), 373–392.

Hirano, M. (1981). Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice. In *Clinical Examination of Voice* (pp. 81–84). London, England: Springer London.

Holmes, N. (1984). *Designer's Guide to Creating Charts and Diagrams*. Broadway, NY: Watson-Guptill Publications.

Hu FB, Sigal RJ, Rich-Edwards JW, & et al. (1999). Walking compared with vigorous physical activity and risk of type 2 diabetes in women: A prospective study. *JAMA*, *282*(15), 1433–1439. http://doi.org/10.1001/jama.282.15.1433

Hunter, E. J. (2012). Teacher response to ambulatory monitoring of voice. *Logopedics Phoniatrics Vocology*, *37*(3), 133–135. http://doi.org/10.3109/14015439.2012.664657

Hunter, E. J., Bottalico, P., Graetzer, S., Leishman, T. W., Berardi, M. L., Eyring, N. G., … Whiting, J. K. (2015). Teachers and Teaching: Speech Production Accommodations Due to Changes in the Acoustic Environment. *Energy Procedia*, *78*, 3102–3107. http://doi.org/10.1016/j.egypro.2015.11.764

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

Hunter, E. J., & Titze, I. R. (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.

Intille, S. S. (2004). Ubiquitous computing technology for just-in-time motivation of behavior change. *Stud Health Technol Inform*, *107*(Pt 2), 1434–7.

Jiang, J., & Stern, J. (2004). Receiver operating characteristic analysis of aerodynamic parameters obtained by airflow interruption: A preliminary report. *Annals of Otology, Rhinology & Laryngology*, *113*(12), 961–966.

Keefe, F. J., Lefebvre, J. C., Kerns, R. D., Rosenberg, R., Beaupre, P., Prochaska, J., … Caldwell, D. S. (2000). Understanding the adoption of arthritis self-management: stages of change profiles among arthritis patients. *Pain*, *87*(3), 303–313.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, *18*(2), 124–132.

Kim, C.-J., Hwang, A.-R., & Yoo, J.-S. (2004). The impact of a stage-matched intervention to promote exercise behavior in participants with type 2 diabetes. *International Journal of Nursing Studies*, *41*(8), 833–841. http://doi.org/10.1016/j.ijnurstu.2004.03.009

Kopf, L. M., Shrivastav, R., & Eddins, D. A. (2013, August). *Isolating the effects of strain on voice quality perception*. Poster presented at the PEVOC, Prague, Czech Republic.

Kopf, L. M., Shrivastav, R., Eddins, D. A., Skowronski, M. D., & Hunter, E. J. (2014). A Comparison of Voice Signal Typing and Pitch Strength. Presented at the American Speech-Language Hearing Association (ASHA) Convention, Orlando, FL.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech, Language and Hearing Research*, *36*(1), 21.

Lee, S. Y., Hwang, H., Hawkins, R., & Pingree, S. (2008). Interplay of Negative Emotion and Health Self-Efficacy on the Use of Health Information and Its Outcomes. *Communication Research*, *35*(3), 358–381. http://doi.org/10.1177/0093650208315962

Lim, B. Y., Shick, A., Harrison, C., & Hudson, S. E. (2011). Pediluma: motivating physical activity through contextual information and social influence. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction* (pp. 173–180). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1935736

Lucas Jr., H. C. (1971). A user-oriented approach to systems design. In *Proceedings of the 1971 26th annual conference* (pp. 325–338). New York, NY, USA: ACM.

Ma, E. P.-M., & Yiu, E. M.-L. (2005). Suitability of acoustic perturbation measures in analysing periodic and nearly periodic voice signals. *Folia Phoniatrica et Logopaedica*, *57*(1), 38–47.

Manson, J. E., Hu, F. B., Rich-Edwards, J. W., Colditz, G. A., Stampfer, M. J., Willett, W. C., … Hennekens, C. H. (1999). A Prospective Study of Walking as Compared with Vigorous Exercise in the Prevention of Coronary Heart Disease in Women. *New England Journal of Medicine*, *341*(9), 650–658. http://doi.org/10.1056/NEJM199908263410904

Marshall, S. J., & Biddle, S. J. (2001). The transtheoretical model of behavior change: a meta-analysis of applications to physical activity and exercise. *Annals of Behavioral Medicine*, *23*(4), 229–246.

Maryn, Y., De Bodt, M., & Roy, N. (2010). The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders. *Journal of Communication Disorders*, *43*(3), 161–174. http://doi.org/10.1016/j.jcomdis.2009.12.004

Mazza, R. (2006). Evaluating Information Visualization Applications with Focus Groups: The CourseVis Experience. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (pp. 1–6). New York, NY, USA: ACM. http://doi.org/10.1145/1168149.1168155

McCabe, D. J., & Titze, I. R. (2002). Chant Therapy For Treating Vocal Fatigue Among Public School TeachersA Preliminary Study. *American Journal of Speech-Language Pathology*, *11*(4), 356–369.

McConnaughy, E. A., Prochaska, J. O., & Velicer, W. F. (1983). Stages of change in psychotherapy: Measurement and sample profiles. *Psychotherapy: Theory, Research & Practice*, *20*(3), 368–375. http://doi.org/http://dx.doi.org.proxy2.cl.msu.edu/10.1037/h0090198

McCrory, E. (2001). Voice therapy outcomes in vocal fold nodules: a retrospective audit. *International Journal of Language & Communication Disorders*, *36*, 19–24.

McNaney, R., Lindsay, S., Ladha, K., Ladha, C., Schofield, G., Ploetz, T., … others. (2011). Cueing swallowing in Parkinson's disease. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 619–622). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1979030

McNaney, R., Othman, M., Richardson, D., Dunphy, P., Amaral, T., Miller, N., … Vines, J. (2016). Speeching: Mobile Crowdsourced Speech Assessment to Support Self-Monitoring and Management for People with Parkinson's (pp. 4464–4476). ACM Press. http://doi.org/10.1145/2858036.2858321

McNaney, R., Poliakov, I., Vines, J., Balaam, M., Zhang, P., & Olivier, P. (2015). LApp: A Speech Loudness Application for People with Parkinson's on Google Glass (pp. 497–500). ACM Press. http://doi.org/10.1145/2702123.2702292

McNaney, R., Vines, J., Roggen, D., Balaam, M., Zhang, P., Poliakov, I., & Olivier, P. (2014). Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's (pp. 2551–2554). ACM Press. http://doi.org/10.1145/2556288.2557092

Misono, S., Banks, K., Gaillard, P., Goding, G. S., & Yueh, B. (2015). The clinical utility of vocal dosimetry for assessing voice rest: Vocal Dosimetry for Voice Rest. *The Laryngoscope*, *125*(1), 171–176. http://doi.org/10.1002/lary.24887

Moore, B. C. J. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, *74*(3), 750. http://doi.org/10.1121/1.389861

Moore, B. C. J., Glasberg, B. R., & Baer, T. (1997). A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. *Journal of the Audio Engineering Society*, *45*(4), 224–240.

Nanjundeswaran, C., Jacobson, B. H., Gartner-Schmidt, J., & Verdolini Abbott, K. (2015). Vocal Fatigue Index (VFI): Development and Validation. *Journal of Voice*, *29*(3). http://doi.org/10.1016/j.jvoice.2014.09.012

Nanjundeswaran, C., Li, N. Y. K., Chan, K. M. K., Wong, R. K. S., Yiu, E. M.-L., & Verdolini-Abbott, K. (2012). Preliminary Data on Prevention and Treatment of Voice Problems in Student Teachers. *Journal of Voice*, *26*(6), 816.e1-816.e12. http://doi.org/10.1016/j.jvoice.2012.04.008

Ohlsson, A.-C., Andersson, E. M., Södersten, M., Simberg, S., Claesson, S., & Barregård, L. (2015). Voice Disorders in Teacher Students—A Prospective Study and a Randomized Controlled Trial. *Journal of Voice*. http://doi.org/10.1016/j.jvoice.2015.09.004

Oinas-Kukkonen, H., & Harjumaa, M. (2008). A systematic framework for designing and evaluating persuasive systems. In *Persuasive technology* (pp. 164–176). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-68504-3_15

Pereira, L. P. de P., Masson, M. L. V., & Carvalho, F. M. (2015). Vocal warm-up and breathing training for teachers: randomized clinical trial. *Revista de Saúde Pública*, *49*, 0–0. http://doi.org/10.1590/S0034-8910.2015049005716

Pinkowski, B. (1993). LPC spectral moments for clustering acoustic transients. *IEEE Transactions on Speech and Audio Processing*, *1*(3), 362–368.

Portone, C., Johns III, M. M., & Hapner, E. R. (2008). A Review of Patient Adherence to the Recommendation for Voice Therapy. *Journal of Voice*, *22*(2), 192–196. http://doi.org/10.1016/j.jvoice.2006.09.009

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of Consulting and Clinical Psychology*, *51*(3), 390–395.

Prochaska, J. O., & DiClemente, C. C. (1984). Self change processes, self efficacy and decisional balance across five stages of smoking cessation. *Progress in Clinical and Biological Research*, *156*, 131–140.

Prochaska, J. O., Norcross, J. C., Fowler, J. L., Follick, M. J., & Abrams, D. B. (1992). Attendance and outcome in a work site weight control program: Processes and stages of change as process and predictor variables. *Addictive Behaviors*, *17*(1), 35–45. http://doi.org/10.1016/0306-4603(92)90051-V

Ramig, L. O., & Verdolini, K. (1998). Treatment efficacy: Voice disorders. *Journal of Speech, Language and Hearing Research*, *41*(1), S101–S106.

Richter, B., Nusseck, M., Spahn, C., & Echternach, M. (2016). Effectiveness of a Voice Training Program for Student Teachers on Vocal Health. *Journal of Voice*, *30*(4), 452–459. http://doi.org/10.1016/j.jvoice.2015.05.005

Rosen, K., Murdoch, B., Folker, J., Vogel, A., Cahill, L., Delatycki, M., & Corben, L. (2010). Automatic method of pause measurement for normal and dysarthric speech. *Clinical Linguistics & Phonetics*, *24*(2), 141–154. http://doi.org/10.3109/02699200903440983

Rossi-Barbosa, L. A., Gama, A. C. C., & Caldeira, A. P. (2015). Association between readiness for behavior change and complaints of vocal problems in teachers. *CoDAS*, *27*(2), 170–177. http://doi.org/10.1590/2317-1782/20152013088

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-Based Clinical Voice Assessment: A Systematic Review. *American Journal of Speech-Language Pathology*, *22*(2), 212. http://doi.org/10.1044/1058-0360(2012/12-0014)

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

Schloneger, M. J. (2014). Assessments of Voice Use, Voice Quality, and Perceived Singing Voice Function Among College/University Singing Students Ages 18-24 Through Simultaneous Ambulatory Monitoring With Accelerometer and Acoustic Transducers. Retrieved from https://kuscholarworks.ku.edu/handle/1808/18407

Schloneger, M. J., & Hunter, E. J. (2016). Assessments of Voice Use and Voice Quality Among College/University Singing Students Ages 18–24 Through Ambulatory Monitoring With a Full Accelerometer Signal. *Journal of Voice*. http://doi.org/10.1016/j.jvoice.2015.12.018

Schneider-Stickler, B., Knell, C., Aichstill, B., & Jocher, W. (2012). Biofeedback on Voice Use in Call Center Agents in Order to Prevent Occupational Voice Disorders. *Journal of Voice*, *26*(1), 51–62. http://doi.org/10.1016/j.jvoice.2010.10.001

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice*, *17*(4), 502–512. http://doi.org/10.1067/S0892-1997(03)00077-8

Shrivastav, R., & Camacho, A. (2010). A Computational Model to Predict Changes in Breathiness Resulting From Variations in Aspiration Noise Level. *Journal of Voice*, *24*(4), 395–405. http://doi.org/10.1016/j.jvoice.2008.12.001

Shrivastav, R., Eddins, D. A., & Anand, S. (2012). Pitch strength of normal and dysphonic voices. *The Journal of the Acoustical Society of America*, *131*, 2261.

Shrivastav, R., Eddins, D. A., & Kopf, L. M. (2014, October). *What is a better descriptor of dysphonic voices- Fundamental frequency or pitch?* Presented at the Fall Voice, San Antonio, Tx.

Shrivastav, R., & Sapienza, C. M. (2003). Objective measures of breathy voice quality obtained using an auditory model. *The Journal of the Acoustical Society of America*, *114*(4), 2217. http://doi.org/10.1121/1.1605414

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language and Hearing Research*, *48*(2), 323.

Slegers, K., Duysburgh, P., & Jacobs, A. (2010). Research methods for involving hearing impaired children in IT innovation. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 781–784). New York, NY, USA: ACM.

Solomon, N. (2008). Vocal fatigue and its relation to vocal hyperfunction. *International Journal of Speech-Language Pathology*, *10*(4), 254–266.

Speyer, R. (2008). Effects of Voice Therapy: A Systematic Review. *Journal of Voice*, *22*(5), 565–580. http://doi.org/10.1016/j.jvoice.2006.10.005

Stemple, J. C. (2000). *Voice Therapy: Clinical Studies* (2nd ed.). Clifton Park, NY: Singular Thomson Learning.

Sundberg, J., & Gauffin, J. (1978). Waveform and spectrum of the glottal voice source. *Speech, Music and Hearing Quarterly Progress and Status Report*, *19*(2–3), 35–50.

Švec, J. G., Popolo, P. S., & Titze, I. R. (2003). Measurement of vocal doses in speech: experimental procedure and signal processing. *Logopedics Phoniatrics Vocology*, *28*(4), 181–192. http://doi.org/10.1080/14015430310018892

Szabo Portela, A., Hammarberg, B., & Södersten, M. (2013). Speaking Fundamental Frequency and Phonation Time during Work and Leisure Time in Vocally Healthy Preschool Teachers Measured with a Voice Accumulator. *Folia Phoniatrica et Logopaedica*, *65*(2), 84–90. http://doi.org/10.1159/000354673

Teixeira, L. C., Rodrigues, A. L. V., Silva, A. F. G. da, Azevedo, R., Gama, A. C. C., & Behlau, M. (2013). The use of the URICA-VOICE questionnaire to identify the stages of adherence to voice treatment. *CoDAS*, *25*(1), 8–15.

Titze, I. R. (2000). *Principles of Voice Production* (2nd ed.). Iowa City, IA: National Center for Voice and Speech.

Titze, I. R. (2012). *Final progress report: Research toward occupational safety in vocalization* (NIH PROJECT NUMBER: R01DC004224) (pp. 1–19). Salt Lake City, UT.

Titze, I. R. (2015). On flow phonation and airflow management. *Journal of Singing*, *72*(1), 57–58.

Titze, I. R., & Hunter, E. J. (2015). Comparison of Vocal Vibration-Dose Measures for Potential-Damage Risk Criteria. *Journal of Speech Language and Hearing Research*, *58*(5), 1425. http://doi.org/10.1044/2015_JSLHR-S-13-0128

Titze, I. R., Hunter, E. J., & Švec, J. G. (2007). Voicing and silence periods in daily and weekly vocalizations of teachers. *The Journal of the Acoustical Society of America*, *121*(1), 469. http://doi.org/10.1121/1.2390676

Titze, I. R., Lemke, J., & Montequin, D. (1997). Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report. *Journal of Voice*, *11*(3), 254–259. http://doi.org/10.1016/S0892-1997(97)80002-1

Titze, I. R., Švec, J. G., & Popolo, P. S. (2003). Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues. *Journal of Speech, Language, and Hearing Research: JSLHR*, *46*(4), 919.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, Connecticut: Graphics Press LLC.

van Leer, E. (2010). *The role of social-cognitive factors in voice therapy adherence and outcomes* (Ph.D.). The University of Wisconsin - Madison, United States -- Wisconsin.

Retrieved from
http://search.proquest.com.proxy2.cl.msu.edu/docview/861307914/abstract?accountid=12
598

van Leer, E., & Connor, N. P. (2010). Patient Perceptions of Voice Therapy Adherence. *Journal of Voice*, *24*(4), 458–469. http://doi.org/10.1016/j.jvoice.2008.12.009

van Leer, E., & Connor, N. P. (2012). Use of Portable Digital Media Players Increases Patient Motivation and Practice in Voice Therapy. *Journal of Voice*, *26*(4), 447–453. http://doi.org/10.1016/j.jvoice.2011.05.006

van Leer, E., Hapner, E. R., & Connor, N. P. (2008). Transtheoretical model of health behavior change applied to voice therapy. *Journal of Voice*, *22*(6), 688–698. http://doi.org/10.1016/j.jvoice.2007.01.011

van Leer, E., Pfister, R. C., & Zhou, X. (2016). An iOS-based Cepstral Peak Prominence Application: Feasibility for Patient Practice of Resonant Voice. *Journal of Voice*. http://doi.org/10.1016/j.jvoice.2015.11.022

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. http://doi.org/10.1044/2015_JSLHR-S-14-0159

Vilkman, E., Lauri, E.-R., Alku, P., Sala, E., & Sihvo, M. (1999). Effects of prolonged oral reading on F 0, SPL, subglottal pressure and amplitude characteristics of glottal flow waveforms. *Journal of Voice*, *13*(2), 303–312.

Villar, A. C. N. W. B., Korn, G. P., & Azevedo, R. R. (2016). Perceptual-auditory and Acoustic Analysis of Air Traffic Controllers' Voices Pre- and Postshift. *Journal of Voice*. http://doi.org/10.1016/j.jvoice.2015.10.021

Warhurst, S., Madill, C., McCabe, P., Ternström, S., Yiu, E., & Heard, R. (2016). Perceptual and Acoustic Analyses of Good Voice Quality in Male Radio Performers. *Journal of Voice*. http://doi.org/10.1016/j.jvoice.2016.05.016

Warhurst, S., McCabe, P., Yiu, E., Heard, R., & Madill, C. (2013). Acoustic Characteristics of Male Commercial and Public Radio Broadcast Voices. *Journal of Voice*, *27*(5), 655.e1-655.e7. http://doi.org/10.1016/j.jvoice.2013.04.012

Werth, K., Voigt, D., Döllinger, M., Eysholdt, U., & Lohscheller, J. (2010). Clinical value of acoustic voice measures: a retrospective study. *European Archives of Oto-Rhino-Laryngology*, *267*(8), 1261–1271. http://doi.org/10.1007/s00405-010-1214-2

Wilcox, N. S., Prochaska, J. O., Velicer, W. F., & DiClemente, C. C. (1985). Subject characteristics as predictors of self-change in smoking. *Addictive Behaviors*, *10*(4), 407–412. http://doi.org/10.1016/0306-4603(85)90037-1

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968.