The effect of speaker-specific information on perceptual voice rating tasks

Cara Sauder

A dissertation
submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Tanya Eadie, Chair

Kristie Spencer

Rita Patel

Susan Joslyn

Program Authorized to Offer Degree:

Speech & Hearing Sciences

University of Washington

Abstract

The effect of speaker-specific information on perceptual voice rating tasks

Cara Sauder

Chair of the Supervisory Committee: Professor Tanya Eadie

Speech & Hearing Sciences

Auditory-perceptual voice assessment and visual-perceptual judgments of videolaryngostroboscopic exams are among the most common and important components of a comprehensive clinical voice assessment. Findings from these perceptual voice assessments are used to determine the presence, severity, and/or nature of a voice disorder. However, perceptual measures are inherently subjective. Error and variability among clinicians are a threat to the validity of these important voice assessment measures. One potential source of variability that has not been systematically investigated relates to what is known about a speaker during perceptual voice assessment. For example, exposure to speaker-specific information (e.g., referring diagnosis, case history, voice quality, etc.) may be controlled in a laboratory setting, but is common when perceptual rating tasks are accomplished in a clinical setting. Understanding the effects of speaker-specific information on perceptual voice assessment measures is important for designing clinical and research protocols and for interpreting results from the existing literature. Because findings from a comprehensive voice assessment are used to form clinical impressions,

determine a clinical diagnosis, and/or make treatment recommendations, it is also important to consider the overall effect of speaker-specific information on clinical decisions.

This document includes a review of the literature (Chapter 1) on the effects of speaker-specific information on auditory- and visual-perceptual videolaryngostroboscopic rating tasks and clinical judgments about voice disorder diagnosis and treatment. Chapter 1 also describes how an existing conceptual model used to explain how an acoustic signal is mapped to auditory-perceptual voice ratings may be extended to visual-perceptual ratings of videolaryngostroboscopy (VLS). Next, three original studies are described in which speaker-specific information that suggests the presence/absence and nature of a voice disorder are provided during auditory-perceptual rating tasks (Chapter 2) and visual-perceptual VLS rating tasks (Chapter 3 and Chapter 4). In Chapter 2, the effect of knowledge of accurate and inaccurate referring diagnosis on auditory-perceptual ratings was investigated. In Chapter 3, the effect of accurate and inaccurate case histories suggesting a particular voice disorder etiology or the absence of a voice disorder (control) on visual-perceptual VLS ratings and other clinical judgments was evaluated. In Chapter 4, the effect of "accurate" auditory cues suggesting different levels of dysphonia severity on visual-perceptual VLS ratings and other clinical judgments was evaluated. The results suggest that the small observed differences in average auditory-perceptual ratings when an accurate referring medical laryngeal diagnosis was present versus absent were unlikely to be clinically meaningful. However, the overall effect of the same referring medical laryngeal diagnoses when they were inaccurate, or inconsistent with the speaker's true medical laryngeal diagnoses, was greater. These differences in perceived severity were also potentially clinically meaningful. These findings are important because referring medical laryngeal diagnoses are commonly inaccurate in a clinical setting. In Chapter 3, there

were also observed differences in the probability of rating VLS parameters as more severe for two of three outcome measures when a preliminary case history (accurate or inaccurate) suggested its presence. Agreement in clinical impressions was increased when case histories were accurate versus inaccurate, and there was also an effect of speaker-specific information on treatment recommendations. Chapter 4 revealed no clinically meaningful effect of the presence of speaker-specific information about dysphonia severity *during* clinicians' evaluation of VLS exams on ratings of VLS parameters. Although some trends in diagnostic codes and treatment recommendations were observed when auditory cues were present versus absent, overall agreement in diagnostic coding and treatment recommendations when auditory cues were absent versus present was high. Findings across these three studies are integrated and summarized in Chapter 5. The results support continued investigations into factors affecting variability in perceptual voice rating tasks and clinical judgments to guide decisions about the best way to perform these important clinical voice assessments in clinical and laboratory settings. Findings are also useful for interpreting and evaluating the external validity of the existing literature.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

46

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisory committee for their expertise, thoughtfulness, and efforts. Tanya Eadie, you provided me with encouragement and an opportunity to pursue an area of research that is fascinating to me. Susan Joslyn, I immensely enjoyed our discussions, and your course in decision-making was excellent. I am so grateful that your expertise in decision-making has influenced the direction of my research. Thank you Rita Patel for all of your wisdom and guidance. Kristie Spencer, thank you for always initiating thoughtful discussions and raising important questions.

I am so grateful for the help and support of Martin Nevdahl and Mara Kapsner-Smith, and all of the members of the Vocal Function Lab. Thank you to the Speech and Hearing Sciences Department for all of the opportunities to teach, gain valuable experience as a research assistant, and travel to conferences to present this work. Also thank you to my fellow doctoral students that provided unwavering support and friendship during this journey.

I would like to acknowledge the expert laryngologists and speech-language pathologists, as well as the graduate student clinicians that agreed to participate in these studies. I am in awe of your generosity. I am incredibly grateful to my family and friends for their encouragement and support over the years. I know the “data collection road trips” across the Midwest and to Canada might not have been the vacations you imagined, but thank you for being so amazing.

ORGANIZATION

This dissertation is organized as three self-contained manuscripts related to investigating the effects of different types of speaker-specific information on perceptual voice assessment rating tasks. Chapter 2, 3, and 4 were written in preparation for publication and can be read independently of the others. There is overlap in background information and references cited among all chapters.

Chapter 1: Sauder, C. "The effect of speaker-specific information on perceptual voice assessment measures and clinical judgments."

Chapter 2: Sauder C, Eadie T. Does the Accuracy of Medical Diagnoses Affect Novice Listeners' Auditory-Perceptual Judgments of Dysphonia Severity? *Journal of Voice*. 2020;34(2):197-207.

Chapter 3: Sauder C, Nevdahl M, Kapsner-Smith M, Merati A, Eadie T. Does the accuracy of case history affect interpretation of videolaryngostroboscopic exams? *The Laryngoscope*. 2020;130(3):718-725.

Chapter 4: Sauder, C. "The effect of auditory information on visual-perceptual ratings of videolaryngostroboscopy and other clinical judgments."

Chapter 5: Sauder, C. "General discussion and summary of findings."

Chapter 1

The effect of speaker-specific information on perceptual voice assessment measures and clinical judgments

Abstract

Research examining sources of variability in auditory-perceptual and visual-perceptual VLS voice rating tasks are commonly conducted in laboratory settings in which raters are provided minimal information about the speaker being evaluated. This differs from clinical settings in which clinicians are exposed to much information about a speaker prior to performing auditory- and visual-perceptual voice assessments. There are many potential sources of variability in perceptual voice rating tasks that might relate to the clinician and the characteristics of the rating task, but few studies have examined task factors that relate to the effect of different types of speaker-specific information on perceptual voice assessment measures. However, several studies in broader fields of medicine have examined the effects of different types of patient information on a variety of clinical assessment tools, including those that rely on perceptual judgments. These studies in broader fields of medicine have also examined the potential effect of patient information on diagnosis and treatment recommendations. Results from these related fields may be used to inform hypotheses and generate a conceptual framework for understanding how patient information may similarly extend to perceptual ratings in voice.

In Chapter 1, research pertaining to the effects of patient/speaker information on perceptual voice rating tasks is reviewed. The evidence suggests that the effect of speaker-specific information on perceptual voice assessment likely requires consideration of several factors. First, the effect of speaker-specific information on perceptual outcomes may be affected by the type and nature of the information itself (e.g., type of information such as diagnosis or risk factors related to a condition, perceived relevance, actual relevance, order of presentation of the information, accuracy of information, etc.). Other factors that require consideration include those related to the rater (e.g., a clinician's experience, bias, profession) and the rating task (e.g.,

severity of the stimuli within a rating set, uni- or multi-dimensionality of the rating parameters, the specific parameters under investigation, and the type of rating tool). Together, these factors are presented in a conceptual model that outlines their effect on auditory-perceptual voice measures. This model is then extended to provide a framework for systematically investigating sources of variability in visual-perceptual VLS assessment. Uniquely, and importantly, this model includes a focus on speaker-specific factors as one type of rating task factor that may affect VLS outcomes. Finally, this framework is presented as an approach for interpreting findings from current clinical and research studies and for helping to design future protocols.

Introduction

Voice disorders are, in part, defined by the American Speech-Language-Hearing Association (ASHA) as inappropriate or abnormal voice production given the "age, gender, cultural background, or geographic location of the individual".[1] This part of ASHA's definition of a voice disorder particularly emphasizes deviations that might cause attention or distress to a listener.[2] In fact, dysphonia, or altered voice quality, is a fundamental sign and/or symptom of a voice disorder. Dysphonia is characterized by altered vocal quality, pitch, loudness, or vocal effort.[1] Kempster et al. stated, "… auditory perceptual measures of voice quality define the presence or absence of a voice disorder clinically."[3(p125)] While auditory-perceptual measures are commonly considered a gold standard for determining the presence of dysphonia, additional clinical voice assessment tools are needed to determine the severity and nature of dysphonia. These tools include other perceptual measures, such as ratings derived from videolaryngostroboscopy (VLS), instrumental voice assessment measures (e.g., acoustics, aerodynamics), and patient-reported measures. All of these tools are necessary for characterizing the nature, severity, etiology, and prognosis of a voice disorder.

ASHA broadly characterizes the nature of voice disorders as organic or functional.[1] Organic voice disorders include structural (e.g., benign or malignant vocal fold lesions, papilloma, scar, etc.) or neurogenic (e.g., spasmodic dysphonia, unilateral vocal fold paralysis) etiologies.[1] Functional voice disorders are those in which one or more of the mechanisms involved in voice production are functioning inappropriately or inefficiently, despite normal appearing physical structures and neurological function. Functional voice disorders are sometimes considered diagnoses of exclusion. One consideration is that individuals can present

with multiple types of voice disorders from within the same category, or from different categories. For example, patients might have a diagnosis of a benign lesion (structural) and a diagnosis of muscle tension dysphonia (functional). Other patients might have a diagnosis of a right vocal fold cyst (structural) with a left reactive lesion (structural).

It is important to accurately determine the nature of a speaker's dysphonia because treatment recommendations often depend on etiology. For example, surgical and/or medical interventions are not typically recommended for speakers with functional voice disorders, but are recommended for speakers with specific structural or neurologic etiologies. If patients are misdiagnosed with a functional disorder when they have a structural or neurological disorder, delays in treatment and loss to follow up might result in poorer patient outcomes.[4] Similarly, when the nature of the dysphonia is functional but patients are misdiagnosed with structural or neurological disorders, they might undergo unnecessary or detrimental surgical and/or medical interventions. Therefore, two important goals of clinical voice assessment are to determine the nature of dysphonia and to make appropriate treatment recommendations.

Otolaryngologists (ENT)s and speech-language pathologists (SLP)s make clinical decisions about the nature of the voice disorder and the best treatment(s) for speakers with dysphonia. ENTs routinely make clinical medical laryngeal diagnoses and determine if medical and/or surgical interventions are appropriate. SLPs assess vocal function and assess the impact of the voice disorder on the patient, determine candidacy for behavioral interventions, and ensure that findings from clinical voice assessment tools are consistent with the medical laryngeal diagnosis.[5] Decision-making, including decisions associated with medical laryngeal diagnosis and/or treatment, typically includes a few key steps.[6] First, decision-making involves the identification of a specific goal related to the severity or nature of the dysphonia, and the most

appropriate treatment and/or the approach for evaluating treatment outcomes. Then, relevant information is collected during a clinical voice assessment. Hypotheses about possible clinical diagnoses and treatments are generated from this information. Clinicians then select one of the alternatives and/or courses of action, and evaluate the outcomes of these decisions.

The intent of Chapter 1 is to provide information about how findings obtained from a comprehensive clinical voice assessment might affect: i) auditory- and visual-perceptual voice assessments; ii) clinical decisions about the presence, severity, or nature of a voice disorder; and iii) clinical decisions about voice treatment. A secondary goal of Chapter 1 is to describe how an existing conceptual model that outlines how an acoustic signal is mapped to auditory-perceptual voice assessment might be extended to visual-perceptual ratings of VLS. The model is presented as a framework for understanding potential sources of variability of visual-perceptual assessments of VLS. Finally, clinical and research implications are discussed.

Comprehensive Clinical Voice Assessment

A comprehensive clinical voice assessment protocol includes a case history, auditory-perceptual voice assessment, patient-reported outcome measures, acoustic and/or aerodynamic analysis of voice, and endoscopic visualization of the larynx and surrounding structures, among other measures.[5,7] An ad hoc committee sponsored by the Voice and Voice Disorders Interest Group (ASHA Special Interest Group 3) was charged with developing a minimal set of recommendations for instrumental voice assessment. The ad hoc committee recommended that both objective and subjective measures of voice be obtained for comprehensive clinical voice disorder assessment.[7] These measures are obtained from a variety of clinical assessment tools (see figure 1). Information from these tools is then combined to determine the presence and

severity of the dysphonia, the etiology of the dysphonia, and the best treatment approach. Information is also used to evaluate the outcomes of previous decisions about diagnosis and treatment, by evaluating treatment outcomes.

Figure 1 Comprehensive voice assessment for clinical diagnosis and treatment

Perceptual Voice Assessment Measures

Auditory-perceptual voice assessment and visual-perceptual assessment of VLS are among the most common and important voice assessment tools used in clinical practice.[5,7,8] However, an existing high quality evidence base to support the validity of these measures to determine the presence, severity, and nature of a voice disorder is limited.[8] The validity of a test means that a test measures what it is intended to measure. It should be sensitive (i.e., able to

identify abnormalities when they exist), and specific (i.e., able to determine that an abnormality is absent when it does not exist), with acceptable overall accuracy (high sensitivity and specificity).[8] Roy and colleagues [8] performed a systematic review of research studies used to establish the validity of clinical voice assessment tools to determine the presence/absence of a voice disorder, the nature of a voice disorder, and/or the severity of the disorder. Studies were evaluated to determine whether they met specific criteria for quality. In this systematic review,[8] very few studies that included auditory-perceptual assessment or visual-perceptual assessment of VLS met the criteria for high quality research studies. This systematic review[8] highlighted a significant gap in the literature, and the need to develop a broader high quality evidence base to support the validity of these important perceptual voice measures for the purposes of diagnostic accuracy.

One intrinsic difficulty in establishing the validity of auditory-perceptual assessment and VLS, is that these clinical tools rely on perceptual judgments. Perceptual measures are inherently subjective in nature, and therefore are prone to error and variability among clinicians who use these measures. Identifying factors associated with error and variability in these perceptual measures is essential for developing reliable and valid clinical protocols, designing research studies, and interpreting findings from the existing literature. Results also have implications for training clinicians.

Auditory-perceptual Voice Assessment

Auditory-perceptual voice assessment is one gold standard for determining the presence of dysphonia clinically[3] and might also routinely be used to corroborate a referring medical laryngeal diagnosis. In a clinical setting, clinicians evaluate and scale many auditory-perceptual

voice parameters (e.g., roughness, breathiness, strain, loudness, and pitch) as well as abnormalities in resonance and/or other aspects of speech.[3] Clinicians determine whether detected abnormalities are present on a consistent or inconsistent basis. Variability in these perceptual ratings amongst clinicians who use these measures can threaten their validity.

Clinical assessment measures must be reliable in order to be valid. There are many factors that affect rater reliability of perceptual judgments. Kreiman and colleagues[9,10] asserted that when listeners make voice ratings, they compare an acoustic stimulus to their internal standards for the voice quality being judged. These internal standards may be affected by many factors, including those related to the listener; the variability in internal standards may also be circumvented by using certain types of rating tasks. Kreiman et al.[9] proposed one conceptual model to explain the many sources of variability in mapping an acoustic voice signal onto auditory-perceptual voice quality ratings (see Figure 2). This model includes listener factors and rating task factors, and the interaction between the two. Factors associated with the rating tasks might relate to the stimuli that are included in the task. For example poor range of voice severity, or the inherent multi-dimensionality of natural speech stimuli might result in poorer rater reliability.[9,11–13] The type of rating scale used to capture the perceptual judgments (e.g., categorical ratings, equal-appearing interval scales, visual analog scales), the rating dimension (e.g., breathiness versus strain), the inclusion of anchors as comparisons, and the type of rating (direct magnitude estimation vs matching) are additional examples of task factors that might affect auditory-perceptual voice assessment.[9,14–16] Voice quality ratings can also be influenced by listener factors, such as experience, sensitivity to the dimension under investigation, bias, or error (e.g., fatigue or inattention). Ratings may also be affected by the interaction between

listener and task factors (e.g., one listener may be particularly sensitive to an attribute such as pitch or breathiness because of previous experience to patients with these qualities).

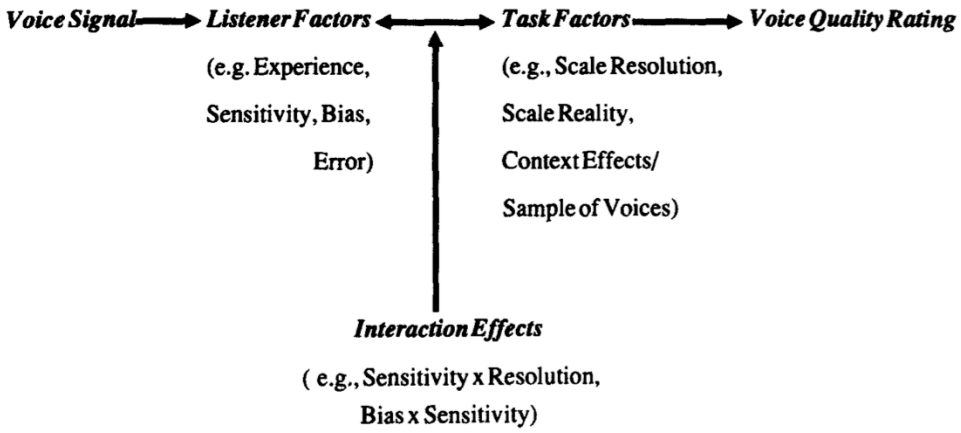


Figure 2 Factors involved in mapping an acoustic signal onto a voice quality rating[9(p32)]

To control for some of the potential sources of variability related to the rating task, and to allow for comparisons between clinicians, the *Consensus Auditory-Perceptual Evaluation of Voice* (CAPE-V) was created.[3,17] This rating tool includes operational definitions of each voice parameter. The CAPE-V uses 100 mm visual analog scales (VAS) for rating severity of deviance for each voice parameter (e.g., overall severity, roughness, breathiness, strain, pitch, loudness) after asking each patient to produce a standard set of voice and speech stimuli. Clinicians also rate any perceived abnormalities in resonance, speech, or other voice characteristics (e.g., tremor, diplophonia).

Although using this standardized method for rating voice and voice quality controls for many sources of variability associated with the rating task, inter- and intra-rater reliability for some auditory-perceptual dimensions, such as strain, remains problematic.[18–20] Additionally, rater reliability is often best near the endpoints of a scale, and worst near the midpoints of the

scale, suggesting that voice signals that are mild or moderate in severity might be most difficult to map to a voice quality rating, compared to voices that are normal or severe.[21,22] Although obtaining adequate rater reliability for some auditory-perceptual voice parameters and/or stimuli continues to be problematic, auditory-perceptual voice assessment remains a gold standard for determining voice disorder status, and is a common practice among voice therapists.[3,5] Compared to auditory-perceptual assessment, there are fewer studies that have examined sources of variability in visual-perceptual judgments of VLS. VLS is another gold standard used to measure the validity of other clinical voice assessment tools. VLS is the most important instrumental clinical voice assessment tool used to evaluate speakers with dysphonia, and it plays a critical role in determining the nature of a voice disorder.[7,8]

Visual-perceptual Ratings of Videolaryngostroboscopy (VLS)

Clinicians use VLS to detect and scale voice abnormalities, make clinical diagnoses, monitor disease progression, and measure treatment outcomes.[5] Clinicians also use information from VLS and other voice assessment tools to make clinical decisions about treatment plans. Clinicians, such as ENTs and/or SLPs, with the appropriate education and training for competency, perform VLS.[23] Some aspects of vocal function (i.e., laryngeal and hypopharyngeal appearance and mobility) can be evaluated endoscopically using a continuous light source. However, vocal fold vibratory characteristics (e.g., vocal fold phase regularity, vibratory amplitude, mucosal wave, phase symmetry, vertical level, vibratory glottal closure pattern, percentage open phase, etc.) require a stroboscopic light source or high speed video (HSV) imaging.[24] Compared to HSV, VLS is more commonly available in clinical settings, and is used to assess many vibratory characteristics.

Recently, an ad hoc committee sponsored by ASHA's Voice and Voice Disorders Interest Group (ASHA Special Interest Group 3) was charged with developing a minimal set of recommendations for instrumental voice assessment for patients evaluated in the United States.[7] The protocol specifies that all individuals with dysphonia be assessed with videoendoscopic measures of laryngeal function, including methods that can be used to evaluate vibratory function, such as VLS.[7] The American Academy of Otolaryngology-Head and Neck Surgery (AAOHNS) also recommends that visualization of the larynx be accomplished when dysphonia persists for 4 weeks, and/or sooner, when there are concerns about serious underlying causes.[25] Endoscopic examination is also recommended prior to additional imaging (e.g., Computerized Tomography), or initiation of reflux or antibiotic medications. The importance of VLS in clinical voice assessment protocols is illustrated by the associated changes in both medical laryngeal diagnoses and treatment recommendations that accompany its implementation.[26–29] These findings and the recommendations from ASHA and AAOHNS suggest that VLS has high relative importance for making clinical decisions about diagnosis and treatment.

Several studies[30–34] have directly examined the validity of VLS for determining the nature of a voice disorder using histopathological analysis, intra-operative findings, and laryngeal electromyography (LEMG) as gold standards for accuracy, and/or measured rater-reliability of VLS visual-perceptual parameters. For example, Paul et al.[32] reported that VLS increased diagnostic accuracy in 8 patients with different types of benign lesions or vocal fold paralysis, compared to clinical history alone. In this study[32] diagnostic accuracy increased from 5% to 68.3% when laryngoscopy and stroboscopy were combined with case history, versus when only case history was presented. This study also supports the notion that clinicians rely on information from VLS to determine diagnosis. However, Paul et al.[32] do not specify which visual-perceptual

parameters clinicians used to make these differential diagnoses. When the VLS parameters used to make medical laryngeal diagnoses are known, rater reliability is commonly unacceptable or unknown,[30,31,35] which is another threat to the validity of VLS. A review of rater reliability and agreement for various VLS parameters highlights the common findings of inconsistencies in rater reliability between, and within, studies. For example, one systematic review[36] found that only 2 of 80 articles that used VLS to measure voice treatment outcomes and met study specific criteria, were considered to have acceptable inter- *and* intra-rater reliability. This systematic review[36] also found that rater reliability for non-vibratory VLS parameters that are not dependent on the stroboscopic effect were better overall. However, others have reported reduced rater reliability for all VLS parameters, irrespective of whether they measure vibratory or non-vibratory features.[37] The variability in clinicians' perceptual judgments about the presence and severity of specific visual-perceptual features might directly contribute to some of the variability in clinical decision-making when findings from VLS are used to determine the presence, severity, or nature of a voice disorder.

Mapping Videolaryngostroboscopic Stimuli onto Visual-perceptual Ratings

Although no known models have been proposed to examine how VLS stimuli are mapped onto visual-perceptual ratings, several sources of variability in VLS ratings have been identified. By adopting and extending Kreiman's[9] model of auditory-perceptual voice assessment to VLS, these sources of variability could also be categorized as those that relate to the rater, the task, and/or their interactions (see figure 3 below). For example, rater factors such as clinician experience and frequency of making VLS ratings, and rating task factors such as severity of VLS exams within a rating set, rating form characteristics (e.g., scale type, metrics, etc.) of VLS

parameters being rated (e.g., mucosal wave versus edge contour), and rater training for the rating task [37–41] are potential sources of variability in making VLS ratings. These factors should be considered and/or controlled in research and clinical protocols. They also need to be addressed when interpreting findings from existing literature.

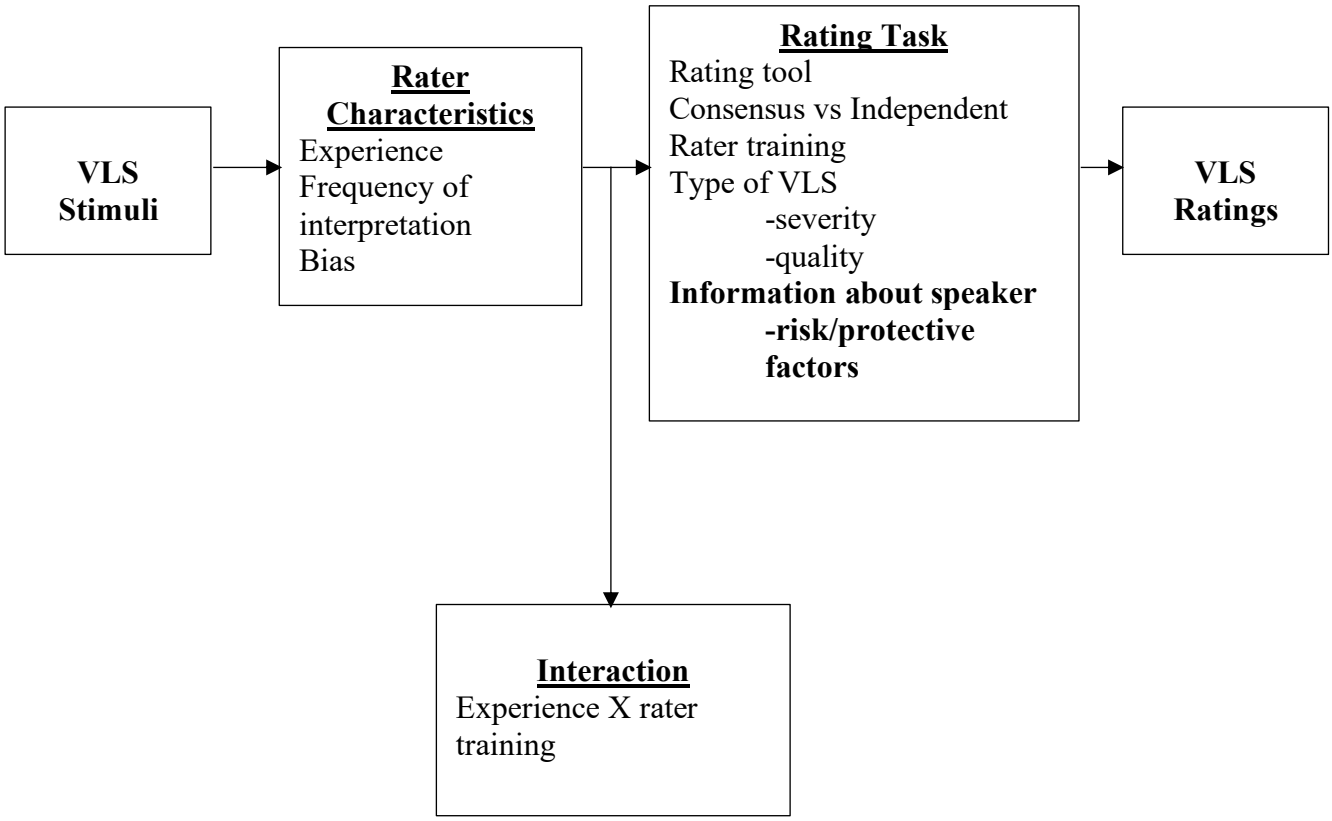


Figure 3 Previously studied sources of variability in visual-perceptual judgments of VLS

One additional rating task factor that needs consideration includes patient/speaker information that is provided to a rater before or during a perceptual rating task. This factor has not been systematically investigated in perceptual voice assessments. It is proposed as a rating

task factor in figure 3 (see bolded section) with regard how it might influence VLS ratings, and may similarly be proposed as a rating task factor in the auditory-perceptual model proposed by Kreiman and colleagues (see figure 3 above). The effect of this information on perceptual voice assessments – both auditory- and visual-perceptual – is the focus of the program of research outlined in this dissertation document and forms the basis for the studies outlined in Chapters 2, 3, and 4.

The Effect of Speaker-specific Information on Perceptual Voice Assessments

While information about a speaker is limited prior to auditory- or visual-perceptual assessments in laboratory settings, auditory- and visual-perceptual judgments in clinical settings are made after case history information is obtained (e.g., referring diagnosis, risk factors) and/or a clinician has been exposed to a speaker's voice quality. Yet, few studies have examined the effects of different types of speaker-specific information on perceptual voice ratings.[42,47] The provision of this additional information might also affect auditory and/or visual-perceptual ratings. For example, based on clinical expertise and knowledge, a case history might result in a preliminary hypothesis about the etiology of voice symptoms. Based on knowledge of the underlying pathophysiology of the etiology, clinicians might then expect certain auditory or visual-perceptual voice features. Consequently, they might be more likely to detect these features and/or perceive them as more severe, and might miss alternative features, or interpret them as less severe. For example, due to the proximity of the recurrent laryngeal nerve to the aorta, cardiac surgery is known a risk factor for vocal fold paralysis/paresis. Exposure to a case history that suggests onset of voice problems following cardiac surgery might result in expectations for increased breathiness and weakness of the voice along with reduced vocal fold mobility and/or

incomplete glottic closure during VLS. These expectations might lead to improved detection and/or the perception of increased severity of breathiness compared to when such a case history is absent.

When auditory- and visual-perceptual judgments are used to make clinical decisions, these differences in detection and perceived severity of perceptual features might also influence clinical diagnosis and treatment recommendations, altering patient outcomes. For example, increased perceived severity of an auditory- or visual-perceptual parameter, such as breathiness or glottal incompetence, might then lead to recommendations for surgical intervention versus behavioral intervention. A similar effect of patient information on auditory and visual-perceptual evaluation of other types of clinical assessments (e.g., radiographs, endoscopy, physical exams, cardiac auscultation) has been observed in broader fields of medicine.[48–52] These findings are commonly attributed to cognitive biases.

While methods used to solve complex problems while minimizing cognitive effort, called heuristics, generally result in fairly good decisions, systematic and predictable deviations from the normative response, called cognitive biases, are also observed during some decision-making tasks.[53,54,55] Although further investigations are needed to determine the neurobiological underpinnings that best explain heuristics and biases, behavioral studies find that cognitive biases are highly predictable.[53] Medical "errors" and variability in perceptual judgments have been attributed to a variety of cognitive biases, but heuristics and biases can have both positive and negative effects on clinical judgements about diagnosis and/or treatment, including those that rely on auditory and visual-perceptual tasks.[48] For example, confirmation bias suggests that examiners might look for, observe, or interpret findings that support a preliminary hypothesis, or impression, and ignore findings that refute a hypothesis.[56] Thus, when patient information is used

to generate a correct preliminary hypothesis, increases in the speed and accuracy of detecting abnormalities, and improved interpretation of these findings for diagnosis and treatment have been observed.[49,50,57] However, when an incorrect preliminary hypothesis is generated, the opposite effect is often observed. Reduced speed and accuracy of detection, and false positives findings are associated with incorrect preliminary hypotheses.

The overall effect of patient information on the detection and interpretation of auditory and/or visual-perceptual judgments is likely a function of the strength of any expectations generated by the information, its accuracy and relevance. However, the effect of patient information on perceptual judgments might also depend on a clinician's stability in these types of judgments. Two previous studies[42,47] examined the effect of speaker-specific information (e.g., referring medical laryngeal diagnosis, risk and protective factors) on auditory and visual-perceptual judgments of VLS, respectively. Findings from these studies form the basis and rationale for the studies outlined in Chapters 2, 3, and 4 of this dissertation, and are discussed next.

Medical Laryngeal Diagnosis and *Auditory-perceptual Voice Assessment*

One study[47] has examined the effect of the presence of an accurate referring medical laryngeal diagnosis on auditory-perceptual ratings of dysphonic speakers. In this study, both novice and experienced clinicians judged roughness and breathiness of dysphonic speakers in two conditions. First, they made judgments of roughness and breathiness of the speakers without any diagnostic information. In the second condition, they evaluated the same speakers in the presence of known diagnostic information (E.g., This person has a vocal fold paralysis). Results showed that novice and experienced clinicians perceived increased severity in the auditory-

perceptual parameter specifically associated with particular diagnostic information. For example, listeners perceived increased roughness, but not breathiness, when they evaluated speakers with known vocal fold lesions. The significant increase in perceived roughness was attributed to the expectations generated by this specific diagnosis (i.e., bilateral lesions that create asymmetry of vocal fold vibration, resulting in increased expectation of roughness). However, there was no effect of these diagnostic labels on unassociated auditory-perceptual parameters. For example, the perceived severity of breathiness, but not roughness increased when listeners rated voices with a known diagnosis of vocal fold paresis/paralysis. The perceived severity of roughness, but not breathiness was increased when listeners rated voices with a known diagnosis of benign vocal fold lesion (e.g., polyp). There was no effect of a referring medical diagnosis of normal larynx on control speakers.

The effect of the diagnostic labels was also greatest for speakers with voices that were mild to moderate in severity, and for ratings of roughness versus breathiness. These findings might reflect a clinician's instability in rating voices of specific severities and perceptual parameters. The effect of referring medical laryngeal diagnoses on severity ratings in this study by Eadie et al.[47] were consistent with a cognitive bias, such as confirmation bias. However, the overall effect of a medical laryngeal diagnosis on auditory-perceptual judgments appeared to depend on factors that might also influence a clinician's reliability in making perceptual ratings.

This study by Eadie et al.[47] provided evidence that a medical laryngeal diagnosis can affect some auditory-perceptual ratings. However, one limitation of the study was that it did not directly examine the effect of the accuracy of the medical laryngeal diagnosis. This is an important consideration for two reasons. First, previous studies[49,50] in broader fields of medicine have observed different effects for inaccurate versus accurate clinical information. Second, the

provision of an inaccurate referring diagnosis is an unfortunate common occurance in clinical voice disorder assessment settings.[26] Consequently, the purpose of the study outlined in Chapter 2 is to investigate the effect of the presence *and* accuracy of the same medical laryngeal diagnoses on auditory-perceptual ratings of perceived roughness and breathiness (see figure 4).

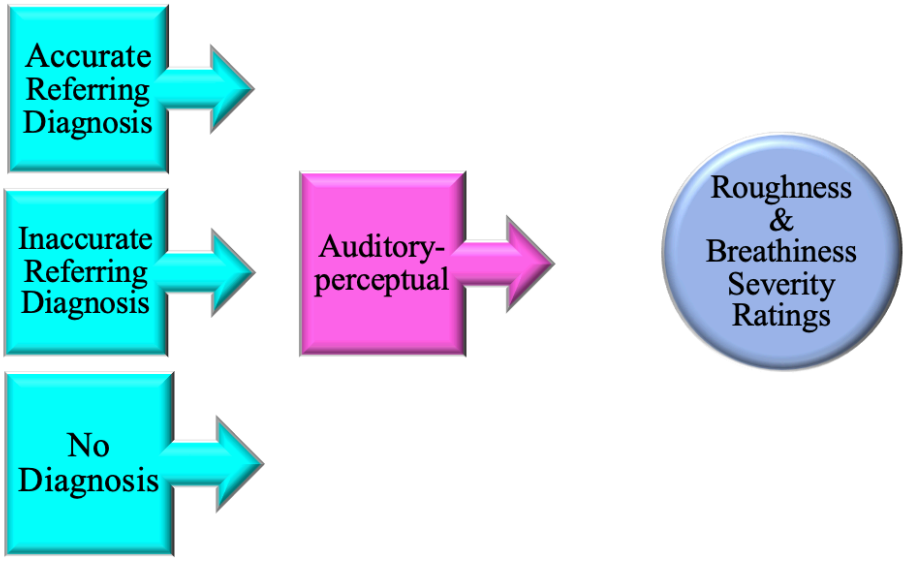


Figure 4 The effect of the presence and accuracy of referring medical diagnoses on roughness and breathiness (basis of study outlined in Chapter 2)

Risk/Protective Factors and Visual-perceptual VLS Ratings

While the effect of speaker-specific information on auditory-perceptual voice assessments is important to consider as it relates to the validity of auditory-perceptual judgments, it is unknown whether speaker-specific information might similarly affect visual-perceptual judgments. It might be hypothesized that VLS as a clinical assessment tool might be particularly susceptible to biases, given the wide range of reported rater reliability of VLS judgments.[36,44,58]

Although no studies have examined the effect of medical laryngeal diagnoses on visual-perceptual VLS ratings, one previous study by Teitler [42] observed systematic differences in perceived VLS severity as a function of speaker-specific information. In that study, clinicians made judgments of VLS parameters in the presence of different types of risk factors for voice disorders (e.g., smoking, dehydration, voice misuse), protective factors for voice disorders (e.g., hydration, voice conservation), or when no case history information was presented. The results of this study showed that when comparing risk and protective factors, VLS severity ratings were in the expected direction. That is, clinicians rated some VLS parameters more severely when the case history indicated risk factors than when they rated the same videos in the presence of protective factors. Similar to the study investigating known laryngeal diagnoses on auditory-perceptual judgments,[47] Teitler [42] also found that the effect of these fictional case histories was greatest for videos characterized as mild to moderate in severity. This finding suggests that when VLS stimuli are mild to moderate, they might be more susceptible to biases.

There were several differences in the effect of control conditions in the visual-perceptual study by Teitler [42] when compared with the auditory-perceptual study by Eadie et al.[47] For example, Teitler [42] observed decreases in perceived abnormality of VLS ratings for VLS exams presented without a case history, compared to either of the fictional case histories (either positive or negative). In contrast, decreases in auditory-perceptual voice severity ratings were not associated with the control condition in Eadie et al.[47] It is possible that this discrepancy between studies could be related to differences in study design, and/or the accuracy of the speaker-specific information presented to raters. Additionally, despite similar levels of clinical experience, the control rater group in Teitler[42] did have "greater regularity of rating VLS exams

in a clinical setting". Unfortunately, Teitler[42] did not complete post hoc analysis of the control group or further discuss this finding.

Overall, comparisons of VLS exams presented with risk or protective factors were consistent with confirmation bias, with the magnitude of the effect dependent on the severity of the VLS stimuli. However, the Teitler[42] study had several limitations. First, because the case histories that were presented with each VLS exam were fictional, the effect of the accuracy, or consistency of the VLS exam presentation with the case history, was not evaluated. Additionally, the author of the study[42] did not investigate whether the observed differences in VLS severity ratings might have clinical implications for diagnosis or treatment. As a result, one purpose of the study outlined in Chapter 3 is to investigate the effect of the accuracy of a case history intended to generate strong expectations for a particular medical laryngeal diagnosis on VLS severity ratings performed by experienced clinicians. In addition, Chapter 3 explores the relationships between VLS exams presented with different case histories, and clinical decisions about diagnosis and treatment.

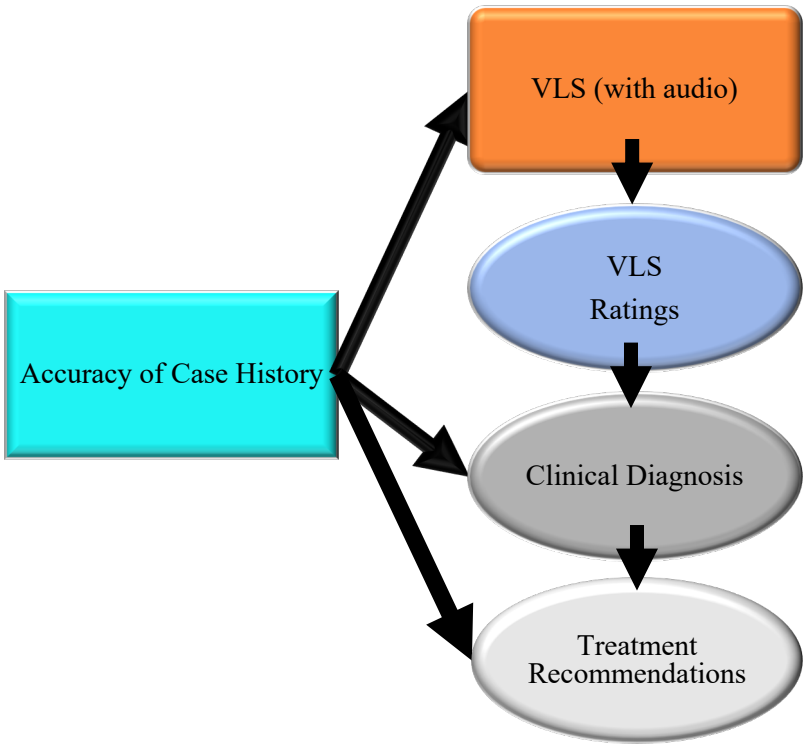


Figure 5 The effect of the accuracy of a case history on VLS ratings, clinical diagnosis and treatment recommendations (basis of study described in Chapter 3)

While Chapters 2 and 3 describe the results of two separate studies investigating the effect of speaker-specific information on auditory- and visual-perceptual voice assessments, it is important to acknowledge that clinicians are exposed to auditory cues *prior* to performing VLS ratings in clinical settings. In addition, auditory cues are also present *during* VLS exams in clinical, and many laboratory settings.[42–44,58] One additional limitation of the previous studies by Teitler[42] and the study outlined in Chapter 3 (Sauder et al.[43]), is that the VLS stimuli in these studies also include auditory information. Auditory-perceptual cues might have provided clinicians with additional speaker-specific information about loudness, pitch, voice quality, and/or dysphonia severity during the VLS rating tasks. This information may have affected the

results of these studies. It is important to examine the effect of these auditory cues because they are commonly present during VLS rating tasks in clinical settings, and they are also present in some laboratory settings. Although auditory information is sometimes excluded from VLS exams to provide comparisons with HSV exams that do not contain auditory information,[7,44] some authors have excluded audio information on the basis that its presence may bias visual-perceptual ratings.[37,38] However, the effect of this auditory information on VLS ratings or clinical judgments about diagnosis and/or treatment is unknown.[37,38,40,46] As a result, the purpose of the study outlined in Chapter 4 is to investigate the effect of the presence of these auditory cues on experienced clinicians' VLS ratings, and on their clinical judgments related to diagnostic coding and treatment.

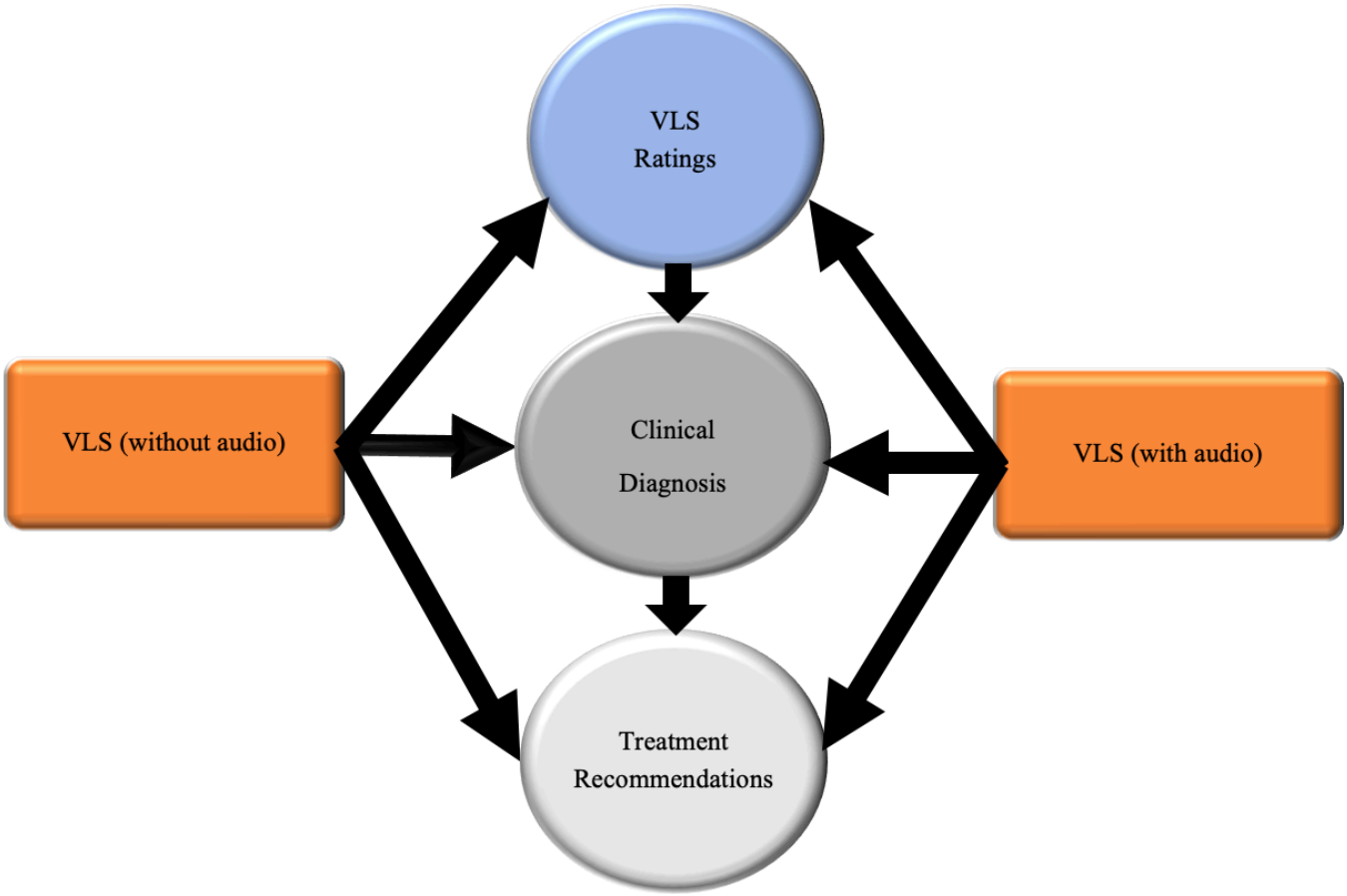


Figure 6 The effect of auditory cues on VLS ratings, clinical diagnosis, and treatment recommendations (basis of study described in Chapter 4)

Together, the studies outlined in chapters 2, 3, and 4 help us to better understand the effects of different types of speaker-specific information on auditory- and visual-perceptual voice assessment measures. In chapter 5, findings from these studies are summarized and integrated. Important factors that might have affected the results of these studies are highlighted. Finally, implications for future research and clinical protocols are discussed.

References

1. ASHA Practice Portal. https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942600&section=Overview.
2. Emerick LL, Van Riper C. *Speech Correction : An Introduction to Speech Pathology and Audiology*. 8th ed. Englewood Cliffs, N.J. : Prentice Hall; 1990.
3. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18(2):124-132.
4. Cohen SM, Kim J, Roy N, Courey M. Delayed otolaryngology referral for voice disorders increases health care costs. *Am J Med*. 2015;128(4):426.e11-18.
5. Behrman A. Common Practices of Voice Therapists in the Evaluation of Patients. *Journal of Voice*. 2005;19(3):454-469.
6. Hausmann D, Zulian C, Battegay E, Zimmerli L. Tracing the decision-making process of physicians with a Decision Process Matrix. *BMC Med Inform Decis Mak*. 2016;16(1):133.
7. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am J Speech Lang Pathol*. 2018;27(3):887-905.
8. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-Based Clinical Voice Assessment: A Systematic Review. 2013;22(2):212-226.
9. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *J Speech Hear Res*. 1993;36(1):21.
10. Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35(3):512-520.
11. Ferrer C. Ferrer, Carlos A., et al. "Collinearity and Sample Coverage Issues in the Objective Measurement of Vocal Quality: The Case of Roughness and Breathiness." Journal of Speech, Language, and Hearing Research, vol. 61, no. 1, 2018, pp. 1–24.
12. Oates J. Auditory-Perceptual Evaluation of Disordered Voice Quality. *Folia Phoniatr Logop*. 2009;61(1):49-56.
13. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America*. 2000;108(4):1867-1876.
14. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20(4):527-544.
15. Misono S, Marmor S, Roy N, Mau T, Cohen SM. Multi-institutional Study of Voice Disorders and Voice Therapy Referral: Report from the CHEER Network. *Otolaryngol Head Neck Surg*. 2016;155(1):33-41.
16. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21(5):576-590.
17. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20(1):14-22.
18. Chan KMK, Yiu EM-L. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45(1):111-126.

19. Eadie TL, Kapsner M, Rosenzweig J, Waugh P, Hillel A, Merati A. The Role of Experience on Judgments of Dysphonia. *Journal of Voice*. 2010;24(5):564-573.
20. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*. 1997;11(1):74-80.21.
21. Kreiman J, Gerratt BR. Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*. 2005;117(4):2201-2211.
22. Eadie TL, Kapsner-Smith M. The Effect of Listener Experience and Anchors on Judgments of Dysphonia. *J Speech Lang Hear Res*. 2011;54(2):430-447.
23. ASHA policy strobe. https://www.asha.org/policy/ks2004-00071/.
24. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J*. 1987;66(7):289-296.
25. Stachler RJ, Francis DO, Schwartz SR, et al. Clinical Practice Guideline: Hoarseness (Dysphonia) (Update) Executive Summary. *Otolaryngol Head Neck Surg*. 2018;158(3):409-426.
26. Cohen SM, Kim J, Roy N, Wilk A, Thomas S, Courey M. Change in diagnosis and treatment following specialty voice evaluation: A national database analysis. *Laryngoscope*. 2015;125(7):1660-1666.
27. Casiano RR, Zaveri V, Lundy DS. Efficacy of Videostroboscopy in the Diagnosis of Voice Disorders. *Otolaryngol Head Neck Surg*. 1992;107(1):95-100.
28. Schwarz K. Swartz, Myers, & Boseley. (2004). Is flexible nasopharyngolaryngoscopy good enough in the diagnosis of the hoarse patient? Otolaryngology - Head and Neck Surgery, 131(2), P139.
29. Woo P, Casper J, Colton R, Brewer D. Diagnosis and Treatment of Persistent Dysphonia After Laryngeal Surgery: A Retrospective Analysis of 62 Patients. *The Laryngoscope*. 1994;104(9):1084-1091.
30. Simpson CB, May LS, Green JK, Eller RL, Jackson CE. Vibratory Asymmetry in Mobile Vocal Folds: Is it Predictive of Vocal Fold Paresis? *Ann Otol Rhinol Laryngol*. 2011;120(4):239-242.
31. Estes C, Sadoughi B, Mauer E, Christos P, Sulica L. Laryngoscopic and stroboscopic signs in the diagnosis of vocal fold paresis: Strobolaryngoscopic Signs for VF Paresis. *The Laryngoscope*. 2017;127(9):2100-2105.
32. Paul BC, Chen S, Sridharan S, Fang Y, Amin MR, Branski RC. Diagnostic accuracy of history, laryngoscopy, and stroboscopy. *Laryngoscope*. 2013;123(1):215-219.
33. Rzepakowska A, Sielska-Badurek E, Osuch-Wojcikiewicz E, Sobol M, Niemczyk K. The predictive value of videostroboscopy in the assessment of premalignant lesions and early glottis cancers. *Otolaryngol Pol*. 2017;71(4):14-18.
34. El-Demerdash A, Fawaz SA, Sabri SM, Sweed A, Rabie H. Sensitivity and specificity of stroboscopy in preoperative differentiation of dysplasia from early invasive glottic carcinoma. *Eur Arch Otorhinolaryngol*. 2015;272(5):1189-1193.
35. Isseroff TF, Parasher AK, Richards A, Sivak M, Woo P. Interrater Reliability in Analysis of Laryngoscopic Features for Unilateral Vocal Fold Paresis. *Journal of Voice*. 2016;30(6):736-740.
36. Bonilha HS, Focht KL, Martin-Harris B. Rater methodology for stroboscopy: a systematic review. *J Voice*. 2015;29(1):101-108.

37. Poburka BJ, Patel RR, Bless DM. Voice-Vibratory Assessment With Laryngeal Imaging (VALI) Form: Reliability of Rating Stroboscopy and High-speed Videoendoscopy. *J Voice*. 2017;31(4):513.e1-513.e14.
38. Rosen CA. Stroboscopy as a research instrument: development of a perceptual evaluation tool. *Laryngoscope*. 2005;115(3):423-428.
39. Nawka T, Konerding U. The interrater reliability of stroboscopy evaluations. *J Voice*. 2012;26(6):812.e1-10.
40. Patel R, Dailey S, Bless D. Comparison of High-Speed Digital Imaging with Stroboscopy for Laryngeal Imaging of Glottal Disorders. *Ann Otol Rhinol Laryngol*. 2008;117(6):413-424.
41. Jones JW, Perryman M, Judge P, et al. Resident Education in Laryngeal Stroboscopy and Perceptual Voice Evaluation: An Assessment. *Journal of Voice*. Published online 2018.
42. Teitler N. Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. *J Voice*. 1995;9(1):95-105.
43. Eadie T, Sroka A, Wright DR, Merati A. Does Knowledge of Medical Diagnosis Bias Auditory-Perceptual Judgments of Dysphonia? *Journal of Voice*. 2011;25(4):420-429.
44. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*. 2016;16(1):138.
45. Shikino K, Ikusaka M, Ohira Y, et al. Influence of predicting the diagnosis from history on the accuracy of physical examination. *Adv Med Educ Pract*. 2015;6:143-148.
46. Leblanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. *Acad Med*. 2002;77(10 Suppl):S67-69.
47. Bytzer P. Information Bias in Endoscopic Assessment. *American Journal of Gastroenterology*. 2007;102(8):1585-1587.
48. Berbaum KS, Brandser EA, Franken EA, Dorfman DD, Caldwell RT, Krupinski EA. Gaze Dwell Times on Acute Trauma Injuries Missed Because of Satisfaction of Search. *Academic Radiology*. 2001;8(4):304-314.
49. Kahneman, D, Slovic, P, & Tversky A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.
50. Cosmides L. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*. 1996;58(1):1-73Cosmides.
51. Elstein AS. Evidence base of clinical diagnosis: Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*. 2002;324(7339):729-732.
52. Pines JM. *Profiles in Patient Safety* : Confirmation Bias in Emergency Medicine. *Academic Emergency Medicine*. 2006;13(1):90-94.
53. Littlefair S, Brennan P, Reed W, Mello-Thoms C. Does Expectation of Abnormality Affect the Search Pattern of Radiologists When Looking for Pulmonary Nodules? *J Digit Imaging*. 2017;30(1):55-62.
54. Poburka B. A Multi-Media, Computer-Based Method for Stroboscopy Rating Training." Journal of Voice : Official Journal of the Voice Foundation., vol. 12, no. 4, 1998, pp. 513–526.
55. Yiu, E., Lau, V., Ma, E., Chan, K., & Barrett, E. (2014). Reliability of laryngostroboscopic evaluation on lesion size and glottal configuration: A revisit. The Laryngoscope, 124(7), 1638-44.

56. Sauder C, Nevdahl M, Kapsner-Smith M, Merati A, Eadie T. Does the accuracy of case history affect interpretation of videolaryngostroboscopic exams? *The Laryngoscope*. 2020;130(3):718-725.
57. Brunings JW, Vanbelle S, Akkermans A, et al. Observer Agreement for Measurements in Videolaryngostroboscopy. *Journal of Voice*. 2018;32(6):756-762.

Chapter 2

Does the Accuracy of Medical Diagnoses Affect Novice Listeners' Auditory-Perceptual Judgments of Dysphonia Severity?

Abstract

Objective/Hypothesis. To determine whether the presence and accuracy of a laryngeal medical diagnosis affects novice listeners' judgments of dysphonia.

Study Design. Prospective, within subjects, modified factorial experimental design.

Methods. Twenty-six speakers with dysphonia and four normophonic speakers provided speech recordings. Forty novice clinicians evaluated speech samples for roughness and breathiness using 100-mm visual analog scales in two conditions. First, speech samples were presented without diagnostic information. In the second condition, 50% of the speech samples were presented with the accurate laryngeal medical diagnosis, while the other 50% of samples were presented with an inaccurate (alternative) diagnosis.

Results. Results showed that judgments of dysphonia were affected by the presence of both accurate and inaccurate diagnoses. As expected, when compared to no known diagnosis, judgments of roughness significantly increased in severity when an accurate diagnostic label of vocal fold lesions was presented. Likewise, in comparison to no known diagnosis, judgments of breathiness trended toward an increase in severity when an accurate diagnostic label of vocal fold paralysis/paresis was presented. Interestingly, increases in perceived severity of dysphonia were also observed with the presentation of inaccurate diagnoses, with the overall effect of inaccurate diagnostic information being greater than accurate diagnoses. Specifically, increases in perceived roughness and breathiness were observed when inaccurate diagnoses included benign vocal fold lesion(s) and vocal fold paralysis/paresis. In contrast, inaccurate diagnostic labels indicating "clear larynx" or diagnoses other than benign vocal fold lesion(s) or paralysis/paresis resulted in decreased perceived roughness and breathiness severity. The

magnitude of the differences in perceived severity between conditions was overall reduced for breathiness compared to roughness.

Conclusions. Sources of bias such as knowledge and accuracy of medical diagnoses should be considered when novice clinicians use auditory-perceptual measures to evaluate dysphonia and measure treatment outcomes.

INTRODUCTION

Auditory-perceptual assessment is a highly valued clinical method used to evaluate individuals with voice disorders. Auditory-perceptual methods are used to determine voice disorder severity, make clinical decisions, and evaluate treatment outcomes. However, auditory-perceptual judgments are subjective and may be susceptible to error.[1] While one source of auditory-perceptual errors may be random, other causes of error may be more systematic. Systematic patterns in judgments are important to identify and understand because, if they result in error, they can potentially be addressed and controlled, thereby reducing any negative effects on clinical and research outcomes.

A systematic deviation from the norm in a perceptual judgment is otherwise known as a cognitive bias.[2] A confirmation bias might occur when information about a patient predisposes a clinician to expect, and consequently seek out an outcome that confirms this expectation. This might lead to an increased sensitivity and perception of that outcome.[3,4] For example, when presented with a referring medical diagnosis of vocal fold paralysis, a clinician's awareness of the underlying physiological abnormalities associated with this diagnosis, such as incomplete glottal closure, might lead to an expectation of breathiness. The expectation that the speaker will be breathy might then lead a clinician to seek out information that confirms this expectation. In other circumstances, a referring diagnosis that indicates a normal laryngeal appearance might result in an expectation that any perceived dysphonia might be less severe. In the behavioral literature, these increases and decreases in perceived severity are described as "overpathologizing" or "minimizing" biases, respectively.[5] It is noteworthy that cognitive biases, such as confirmation biases, may have both positive and negative effects on detecting abnormalities.[6,7] Because clinical information such as a referring medical diagnosis is often

known before voice evaluations are performed, it is important to understand how this could result in a systematic bias in outcomes. These results have implications for determining the validity of perceptual measures used during clinical voice assessment.[8,10]

Bias in clinical voice assessment

Very few studies have examined how confirmation biases might affect perceptual judgments of voice disorders. One prior study examined how knowledge of a fictional patient history affected clinicians' judgments of visual-perceptual measures.[9] In that study, clinicians made judgments of videolaryngostroboscopic parameters in the presence of a positive case history (e.g., a patient with good vocal hygiene, treatment adherence, etc.) or in the presence of a negative case history (e.g., a patient who smokes cigarettes, history of phonotraumatic behaviors, etc.). The study revealed that perceptions of some videolaryngostroboscopic parameters became more severe when clinicians were presented with the negative, as opposed to the positive, case history. In other words, clinicians appeared to be susceptible to a confirmation bias for some visual-perceptual judgments of vocal function.

How diagnostic information affects a clinician's auditory-perceptual evaluation of voice quality also has been examined in two prior studies.[10,11] In the first study, listeners judged speakers as significantly more hypernasal when they knew that speakers had a diagnosis of cleft palate.[11] In the second study, novice and experienced listeners similarly showed confirmation biases when they made auditory-perceptual judgments of dysphonic speakers. In that study, clinicians judged roughness and breathiness of dysphonic speakers in two conditions. First, they made judgments without any diagnostic information. Second, they evaluated the same speakers in the presence of diagnostic information. The results showed that overall, the severity of

listeners' perceptions of voice quality increased when diagnoses were known, with an important caveat. Increases in perceived severity of dysphonia were only observed in the expected dimension (roughness/breathiness) when they were presented in conjunction with particular diagnoses. Listeners perceived increased roughness, but not breathiness, when they evaluated speakers with vocal fold lesions in the presence of this type of diagnostic information. Likewise, listeners' perceptions of breathiness, but not roughness, tended to increase when they made judgments in the presence of known diagnoses of vocal fold paralysis/paresis. Interestingly, no perceived changes in roughness or breathiness were observed in a group of speakers with a variety of laryngeal diagnoses, such as muscle tension dysphonia. It was presumed that knowledge of these types of diagnoses did not increase a clinician's expectation of any predictable voice quality, and, therefore, this category of speakers acted as a control group.

Results from these two studies[10,11] showed that clinicians may be susceptible to a confirmation bias when performing auditory-perceptual judgments of voice quality. However, one limitation to these studies was that listeners were provided only accurate diagnoses. Whether differences in voice severity judgments would occur when these diagnostic labels were inaccurate was not investigated. This is an important consideration because a recent study[12] showed that half of all patients referred to a specialty voice clinic from an outpatient otolaryngology practice had a change in laryngeal diagnosis following specialized voice assessment. These results indicate that it is possible (and perhaps even common) that the referring laryngeal medical diagnoses clinicians encounter prior to voice assessment are inaccurate or nonspecific. Thus, it is critical to consider how the accuracy of such information might also influence auditory-perceptual measures.

Effect of accuracy of diagnostic information on perceptual judgments: implications for voice assessment

While no studies have examined the effect of inaccurate diagnostic information on clinicians' judgments of voice quality, results from the broader medical literature may be used to generate hypotheses related to voice assessment. Overall, studies examining clinical assessment tools that rely on both auditory- and visual-perceptual judgments support the hypothesis that accurate clinical information results in improved diagnostic accuracy, while inaccurate clinical information substantially reduces diagnostic accuracy.[6-8] For example, Shikino et al[8] examined the relationship between preliminary diagnostic information and diagnostic impressions following cervical auscultation of simulated heart murmurs performed by medical students. In this study, students were provided accurate, inaccurate, or no diagnostic information. Diagnostic accuracy was improved when preliminary information was present versus absent, but this difference was not statistically significant. When correct versus inaccurate clinical information was provided, an accurate diagnosis was generated in 87% versus 30% of trials, demonstrating the strong effect of accuracy of information on outcomes.

The accuracy of clinical information has been shown to affect detection and interpretation of visual-perceptual features that lead to different diagnoses in two previous studies.[6,7] In the first study, medical students, residents, and cardiologists identified and interpreted visual electrocardiogram features.[6] Prior to making their judgments, they received no history, a history that suggested the correct diagnosis, or a history that suggested a plausible, but incorrect diagnosis. The results showed that all clinicians were more likely to make judgments that were consistent with the preliminary diagnosis, irrespective of the history's accuracy. These results are also consistent with a confirmation bias.

In a second study, Leblanc et al[7] provided medical students and family medicine residents a preliminary referring diagnosis and eight photographs considered to be highly representative of a particular recognizable diagnosis. When an accurate diagnosis was presented, more relevant visual-perceptual features were identified and diagnostic accuracy was improved. However, when examiners were given inaccurate diagnostic information, a greater percentage of normal features were perceived as abnormal, and fewer abnormal features pertinent to a correct diagnosis were identified. In these cases, diagnostic accuracy was reduced. Similar to the study by Hatala et al,[6] these findings also suggested that irrespective of clinical experience, subjects were susceptible to confirmation biases. When the referring diagnosis was incorrect, these biases resulted in more errors in visual-perceptual judgments and diagnostic inaccuracies. Therefore, the effect of the accuracy of preliminary diagnostic information on auditory-perceptual voice features is an important focus of the present study.

Purpose of study

Auditory-perceptual judgments of voice quality contribute to diagnostic impressions, are used to determine dysphonia severity, measure treatment outcomes, and make clinical decisions. These auditory-perceptual judgments often are made in the presence of information such as the laryngeal medical diagnosis, treatment status, or other clinical signs and symptoms. Yet, the referring laryngeal medical diagnosis might be accurate or inaccurate.[12] While the effect of diagnostic information has been shown to affect auditory-perceptual judgments of dysphonia[10] and its accuracy has been shown to affect perceptual judgments in other fields of medicine,[68] no studies have examined the effect of the accuracy of diagnostic information on auditory-perceptual judgments in voice disorders. Therefore, the purpose of this study was to determine whether the presence and accuracy of a laryngeal medical diagnosis affects judgments of

dysphonia. Results have implications for determining the validity of auditory-perceptual measures used in both research and clinical practice.

METHODS

Speaker samples

Speech samples from 26 speakers with voice disorders (13 males and 13 females) and four speech samples from speakers with no known history of voice impairment and without dysphonia were selected from a database, as described in the study by Eadie et al.[10] Speech samples from those with voice disorders were obtained during a standard evaluation protocol for patients with voice complaints at the Laryngology Clinic at the University of Washington Medical Center. Speakers ranged in age from 25 to 84 years of age (mean = 52.5 years), and had been previously diagnosed by an experienced laryngologist based on a comprehensive voice evaluation that included a case history and videolaryngostroboscopy.

Speech samples selected for inclusion in this study were required to meet a number of criteria. First, speech samples needed to be distributed across the severity spectrum for both breathiness and roughness. The severity of the samples was characterized by the average auditory-perceptual judgments of eight speech-language pathologists with more than 6 years of experience working with individuals with dysphonia. The samples were judged without diagnostic information in the prior study.[10] Speech samples were judged for roughness and breathiness using the Consensus Auditory-Perceptual Evaluation of Voice.[13] Each of the voice dimensions of roughness and breathiness were rated on a 100-mm visual analog scale (VAS), which ranges from 0 (normal) to 100 (very severe). Speech samples were categorized by severity of both roughness and breathiness using the following severity ranges: normal 0-9 mm, mild 10-29 mm, moderate 30-59 mm, and severe 60-100 mm.[14] The included speech samples

demonstrated the following breathiness distributions: normal (43.3%), mild (23.3%), moderate (20%), and severe (13.3%). Roughness also was similarly distributed: normal (30%), mild (33.3%), moderate (23.3%), and severe (13.3%).

A second criterion for selecting speech samples for inclusion in this study related to the speakers' primary laryngeal diagnoses. There was a concerted effort to identify relatively equal numbers across three broad diagnostic categories,[10] which included the following: (a) unilateral or bilateral mass lesions (e.g., unilateral or bilateral mass lesions, including cysts, vocal nodules, papilloma, Reinke's edema, and reflux laryngitis; n = 11); (b) unilateral or bilateral vocal fold paralysis or paresis (n = 9); and (c) control (e.g., muscle tension dysphonia, postoperative radiation effects, subglottic stenosis, and normal larynges; n = 10). Speakers in these categories were included so that hypotheses related to diagnoses and confirmation bias could be tested. For example, consistent with the previous study,[10] increases in perceived roughness severity were expected when speakers with vocal fold lesions were judged in conjunction with known (and accurate) diagnostic labels. In contrast, increases in perceived breathiness severity were expected when speech samples were evaluated in the presence of known (and accurate) diagnoses of vocal fold paresis/ paralysis. Finally, no consistent effects were hypothesized for dysphonic and normophonic speakers in the control group with known (and accurate) diagnostic labels. Using this approach, one condition of this study replicated the previous study related to the effect of the presence of accurate diagnostic information on listeners' judgments of dysphonia. The reader is referred to Eadie et al.[10] for further demographic information related to the speech samples.

Listeners

Forty graduate speech-language pathology students (females = 37; males = 3) were recruited as novice clinician listeners. Novice clinicians were selected for two main reasons. First, like experienced clinicians, novice clinicians have been shown to be susceptible to biases in making perceptual judgments, including those related to voice quality.[6,7,10] Second, if biases among novice listeners are better understood, the results could provide better directions for training this group, as well as identifying possible methods for controlling any observed systematic biases associated with error.

All listeners who participated in this study had completed at least one graduate level course in voice disorders. Coursework included practice with auditory-perceptual measures of voice and experience using VAS, knowledge of voice disorder etiologies and pathologies, and how these etiologies could affect vocal physiology and ultimately voice quality. Most students had at least one practical experience with individuals with voice disorders. Listeners were native English speakers without hearing disorders that passed 20dB sound pressure level (SPL) hearing screening tests bilaterally for frequencies between 250 and 4000 Hz. All listeners were paid and the procedures were approved by the University of Washington Human Subjects Committee.

Speech recordings and preparation of perceptual program

All speech samples were recorded at 44.1 kHz, low-pass filtered at 10 kHz, and downsampled to 25 kHz. The second sentence of the Fairbanks'[15] Rainbow passage, "The rainbow is a division of bright light into many beautiful colors," was extracted using sound-editing software. Samples were normalized for peak intensity and entered into a custom-made computer program that generated randomized speaker order, presented diagnostic information (accurate or inaccurate), presented rating scales, and recorded responses.

Prior to the listening procedure, the experimental program was set up as follows. In the first condition, the perceptual program was designed to present speech samples to listeners without diagnostic information. In the second condition, the program was designed to present speech samples to listeners along with an accurate (50% of samples) or an inaccurate (50% of samples) diagnostic label. To ensure that the inaccurate (alternative) diagnosis was plausible, a number of procedures were followed. First, a composite dysphonic severity score was created for each of the speech samples by averaging the breathiness and roughness ratings from the eight experienced listeners. Next, the severity of each sample was sorted within each diagnostic category (lesions, paralysis/paresis, and control). Finally, an inaccurate (alternative) diagnosis was then selected from a speech sample in an alternative speech sample category that was similar in severity. Using this method, the inaccurate (alternative) diagnoses that were provided in the second condition were always presented with a speech sample that was similar in dysphonic severity, ensuring that the presented diagnosis was plausible for the listener.[7]

Listening procedure

Forty listeners participated in similar listening protocols. Listeners were first provided with simultaneous written and aural definitions of roughness and breathiness. Roughness was defined as the "perceived irregularity in the voicing source," and breathiness was defined as the "audible air escape in the voice," consistent with definitions found in the Consensus Auditory-Perceptual Evaluation of Voice.[13] All listeners were familiarized with 100 mm VAS, with the left endpoints labeled as "normal, no roughness" or “normal, no breathiness,” and the right endpoints marked as “severely rough” or “severely breathy.”

Listeners were then asked to make judgments of roughness or breathiness using 100-mm VAS in the first condition (no diagnosis). This “no diagnostic” condition was performed first for

all listeners to control for learning effects (i.e., to prevent listeners from remembering particular diagnoses for particular speakers). The order in which the perceptual dimension (roughness or breathiness) was rated was counterbalanced, so that 50% of listeners rated the breathiness dimension first, and 50% rated the roughness dimension first, to control for sequencing effects.

After rating both roughness and breathiness in the first condition (no diagnosis), listeners then participated in the second rating condition (diagnostic condition). In this condition, the primary (accurate) diagnosis was presented to listeners in tandem with 50% of the speech samples (n = 15 speech samples per dimension), and an inaccurate (alternative) diagnosis from another speaker category (lesion(s), paralysis/paresis, and control) was presented along with the other 50% of the speech samples (n = 15 speech samples per dimension). This design (50:50) was implemented to avoid having a design in which 100% of the samples were presented all with inaccurate diagnoses, thereby reducing the plausibility of the experiment. In the diagnostic condition, listeners were provided a diagnostic label on the computer screen in tandem with presentation of the speech sample. Listeners were informed that this was the diagnosis of the speaker (e.g., "This speaker has a unilateral vocal fold paralysis"), but no instructions were provided about how to use this information when making their auditory-perceptual ratings. The order of the perceptual dimensions (roughness and breathiness) within this second condition was also counterbalanced across listeners. Thus, in total, listeners participated in four consecutive rating session blocks: roughness (no diagnosis), breathiness (no diagnosis), roughness (diagnosis—50% accurate and 50% inaccurate labels), and breathiness (diagnosis—50% accurate and 50% inaccurate labels).

Stimuli were presented at a comfortable loudness level through headphones (Samson RH600). All ratings were performed after a single presentation of each speech stimulus,

presented in random order within each block. Listeners controlled the rate of presentation of stimuli and took breaks to minimize fatigue. The entire listening experiment lasted approximately 30 minutes and ratings were obtained during a single listening session. Twenty percent of samples (n = 6) in each condition were repeated to a subgroup of 20 listeners to determine intra-rater reliability.

Statistical analysis

IBM SPSS Statistics software version 24 was used to complete statistical analyses.[16] Two linear mixed effects regression (LMER) models were used to examine how the multiple effects of speaker sample category (i.e., the true diagnosis of the speaker), dysphonic severity, and diagnostic label category (i.e., the diagnostic label provided to the listener) on perceived severity of roughness and breathiness. LMER models were chosen for this analysis because they take into account the random effects associated with individual listeners across within subject measures.[17,18]

The overall significance of each speaker category and diagnostic label category were assessed using Sattherwaite's approximation to the denominator or degrees of freedom. The LMER model includes an interaction between the speaker category and the diagnostic label category presented to listeners. This interaction provides specific information about the effects of no diagnostic label, accurate diagnostic labels, or inaccurate diagnostic labels from two alternative speech sample categories, on voice severity for speakers within each category (lesion(s), paralysis, and control). Residual diagnostics were used to ensure model fit.[18]

Least squares group means with approximate standard errors were then used to provide information about the contrasts between no diagnostic label and an accurate diagnostic label as

well as contrasts between accurate and inaccurate diagnostic labels from two alternative speaker categories. These contrasts were of interest given the hypotheses that the type of inaccurate diagnostic label might be important predictors of breathiness or roughness severity. In total, analyses were performed on 1200 ratings of roughness and 1200 ratings of breathiness in the no diagnosis condition (n = 30 speech samples X 40 listeners = 1200 ratings per dimension). Similarly, analyses were based on 1200 ratings of both roughness and breathiness in the diagnostic condition. However, given that half of the ratings were performed for inaccurate (alternative) labels in the diagnostic condition, analyses were based on 600 ratings for accurate and 600 inaccurate diagnoses for both breathiness and roughness. Further, because there were two possible inaccurate diagnoses (e.g., speaker with a true diagnosis of bilateral lesions evaluated in the presence of an inaccurate label of paralysis/paresis or functional/clear larynx [i.e., the control condition]), the average ratings in these conditions were based on 300 observations.

Reliability

Pearson's product correlations and their 95% confidence intervals (CI) were calculated to determine intra-rater reliability for repeated ratings of the same speech samples under the same conditions. The average intra-rater reliability correlations for ratings of breathiness across conditions were $r = 0.979$ (95% CI = 0.95-0.99); the average correlations for ratings of roughness across conditions were $r = 0.525$ (95% CI = 0.19-0.79). A measure of intra-rater agreement also was calculated to determine the degree to which two judgments of the same speech sample agreed within 10 mm on the 100 mm VAS[19] when speakers rated speech samples in the same conditions. The average intra-rater agreement values and 95% CI for ratings of breathiness were

83% (95% CI = 0.76-0.90) across conditions, and 56% (CI = 0.47-0.65) for ratings of roughness across conditions.

Intraclass correlation coefficients (ICCs) and their 95% CI were calculated using a generalization of ICC (2, 1) to assess inter-rater reliability of a single measure ($k = 1$), absolute agreement, two-way random effects model. The model included both the effect of rater and the speech sample and assumed both were drawn randomly from larger populations. ICCs for roughness judgments across all conditions were 0.60 (95% CI = 0.58-0.62) for roughness, and 0.72 (95% CI = 0.70-0.74) for breathiness. Measures of interrater agreement (%) and 95% CI for each of the rating conditions was calculated to determine to what extent listeners agreed with the average severity ratings within each condition. Ratings within 10 mm of the average severity rating on the 100-mm VAS were considered to agree. Mean percent agreement was 53% (CI = 0.51-0.55) for roughness and 61% (CI = 0.59-0.64) for breathiness, consistent with previous auditory-perceptual investigations.[10]

RESULTS

Descriptive statistics

The average judgments of roughness for speaker samples within each category (i.e., the true diagnosis of the speaker) were calculated across all of the listening conditions (no diagnostic label [no Dx], accurate diagnostic [Dx] label, inaccurate diagnostic [Dx] label lesion(s), inaccurate diagnostic [Dx] label paralysis/paresis, and inaccurate diagnostic [Dx] label control). The means and standard errors of the mean (SEMs) are presented in Table 1.

TABLE 1.

Means and SEMs for Average Ratings of Roughness Across Listening Conditions

			Inaccurate Dx		
			Lesion(s)	Paralysis/Paresis	Control
Speaker Category	No Dx M (SEM)	Accurate Dx M (SEM)	M (SEM)	M (SEM)	M (SEM)
Lesion(s)	29.8 (1.6)	33.5 (1.8)	NA	36.9 (2.3)	29.6 (2.2)
Paralysis/paresis	40.4 (1.6)	39.9 (2.0)	45.3 (2.1)	NA	33.3 (2.5)
Control	37.0 (1.6)	35.6 (1.9)	41.6 (2.2)	39.9 (2.3)	NA

Abbreviations: M, mean; SEM, standard error of the mean; Dx, diagnosis; NA, not applicable.

Ratings in the no label (no Dx) condition are based on 1200 total judgments; ratings of presented with accurate labels are based on 600 judgments, and ratings within each of the inaccurate label conditions are based on 300 judgments. Note that within each of the inaccurate label conditions, there are only two possible alternatives; the third condition represents the actual/true diagnosis.

Similarly, the average judgments of breathiness for speaker samples within each category (i.e., the true diagnosis of the speaker) also were calculated across all of the rating conditions. The means and SEMs are presented in Table 2 for descriptive purposes.

TABLE 2.

Means and SEMs for Average Ratings of Breathiness Across Listening Conditions

			Inaccurate Dx		
Speaker Category	No Dx M (SEM)	Accurate Dx M (SEM)	Lesion(s) M (SEM)	Paralysis/Paresis M (SEM)	Control M (SEM)
Lesion(s)	32.3 (1.4)	34.3 (1.6)	NA	32.9 (2.0)	33.9 (1.9)
Paralysis/paresis	39.2 (1.5)	42.3 (1.7)	40.0 (1.9)	NA	34.4 (2.1)
Control	32.3 (1.4)	33.6 (1.7)	31.8 (1.9)	37.3 (2.0)	NA

Abbreviations: M, mean; SEM, standard error of the mean; Dx, diagnosis; NA, not applicable.

Roughness model

To answer the experimental questions related to judgments of roughness, an LMER model was used. The three fixed effects were speaker sample category (i.e., the true diagnosis of the speaker), dysphonic severity, and diagnostic label category (i.e., the diagnostic label provided to the listener). The interaction between speaker sample category and diagnostic label category was also included in the model and provided information about the effect of diagnostic labels on severity ratings for different speaker groups. There were no significant departures from the modeling assumptions such that when the variables in the model were accounted for, residuals were normally distributed.

Results from the model for roughness showed that the speaker category, $F(2, 2346) = 25.03$, $P < 0.0001$, diagnostic label category, $F(3, 2346)= 9.04$, $P < 0.0001$, and expert ratings of roughness severity, $F(3, 2386) = 878.13$, $P < 0.0001$ were significant predictors of roughness ratings from novice listeners in this study. Because the severity ratings from expert listeners were expected to significantly relate to ratings from novice listeners (supporting the validity of the selected speech samples), they will not be discussed further.

The main effects of speaker category showed that across all diagnostic conditions (no diagnosis, accurate, and inaccurate), speakers with paralysis/paresis were judged to be the most rough ($M = 39.75$; $SEM = 1.54$). Speakers with lesions ($M = 32.46$; $SEM = 1.53$) were significantly less rough than both speakers with paralysis/ paresis and speakers in the control group ($M = 38.53$; $SEM = 1.53$) across conditions, $P < 0.0001$.

The main effect of diagnostic label category does not consider the accuracy of this label across the conditions, but is a comparison between ratings performed with no label versus presence of any kind of label (whether accurate or not). Compared to the no diagnostic label condition ($M = 35.69$; $SEM = 1.42$), roughness was significantly increased for speakers presented with a diagnostic label indicating lesion(s) ($M = 39.31$; $SEM = 1.65$) and paralysis ($M = 39.15$; SEM = 1.64), $P = 0.001$ and P = 0.002, respectively. In comparison to the no diagnosis condition, speech samples labeled control ($M = 33.43$; standard deviation [SD] = 1.65) were judged to be significantly less rough, $P = 0.05$. Across all conditions, speakers presented with diagnostic labels of lesion(s) or paralysis/paresis were judged to be significantly more rough than speakers presented with diagnoses from the control group, $P < 0.0001$. The interaction between speaker category and listening condition significantly contributed to the model for roughness severity, $F(6, 2346) = 2.42$, $P = 0.03$. These post hoc contrasts provided information about the

effect of the diagnostic labels on severity ratings for different speaker groups. First, ratings of roughness were compared when speakers were judged without a known diagnosis and when they were judged with an accurate diagnostic label. For speakers with vocal fold lesions, roughness severity significantly increased, on average, 4 mm (95% CI = 0.6-6.6) when ratings were obtained in the presence of an accurate diagnostic label in comparison to no diagnostic label, P = 0.02. No differences in roughness were found for speakers from the control or paralysis/paresis speaker category when listeners were provided with no diagnosis versus an accurate diagnostic label.

Ratings for each speaker category (i.e., the speaker's true diagnosis) were then examined to determine the specific effects of the presented diagnostic labels and their accuracy (or inaccuracy) on perceived roughness. For speakers with lesion(s) or paralysis/paresis, inaccurate diagnostic labels indicating either paralysis/paresis or lesion(s) resulted in increased perceived roughness. For both speaker categories, inaccurate diagnostic labels from the control group resulted in decreased severity of roughness compared to accurate diagnoses. Significant differences in average roughness severity between groups of speakers presented to listeners with these different inaccurate diagnostic labels also were observed. For example, on average, there was a 12 mm (95% CI = 6.0-17.1) difference in perceived roughness when subsets of speakers with paralysis/paresis were provided with an inaccurate diagnosis of lesion(s) versus a diagnosis mismatched from the control group, $P < 0.0001$. Likewise, speakers with lesion(s) differed, on average, by 7 mm (95% CI = 3.1-11.2), $P = 0.001$ when a diagnosis of paralysis/paresis was presented in comparison to a diagnosis mismatched from the control group. These contrasts are presented in Figure 1.

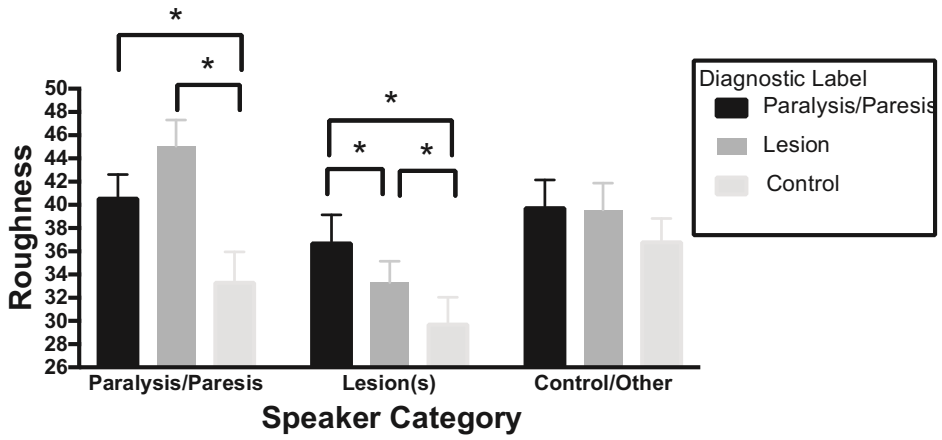


FIGURE 1. Roughness severity by speaker category and diagnostic label.

Breathiness model

To answer the experimental questions related to judgments of breathiness, a second LMER model was used. The three fixed effects were speaker category (i.e., the true diagnosis of the speaker), dysphonic severity, and diagnostic label category (i.e., the diagnostic label provided to the listener). The interaction between speaker sample category and diagnostic label category were also included in the model and provided information about the effect of diagnostic labels on severity ratings for different speaker groups. There were no significant departures from the modeling assumptions such that when the variables in the model were accounted for, residuals were normally distributed.

Results from the LMER model for breathiness showed that speaker sample category, $F(2, 2346) = 18.30$, $P < 0.0001$, diagnostic label category, $F(3, 2346) = 3.94$, $P = 0.008$, and expert ratings of breathiness severity, $F(3, 2346) = 2001.11$, $P < 0.0001$ were significant predictors of breathiness ratings from novice listeners. To more closely examine the main effect of the diagnostic labels (none, accurate, and inaccurate diagnostic labels) on novice listeners' breathiness severity ratings, differences in least significant means were examined. Speech

samples from the paralysis/paresis speech sample category were perceived as significantly more breathy (M = 39.03; SD = 1.38) than speech samples in the lesion(s) category (M = 33.30; SD = 1.33) or control group (M = 33.82; SD = 1.35) irrespective of what diagnostic label was provided (none, accurate, and inaccurate). On average, breathiness severity ratings were increased by 6 mm (95% CI 3.80-7.66) for speakers with paralysis or paresis compared to speakers with lesion(s) and were increased by 5 mm (95% CI 3.10-7.31) compared to speakers in the control group, $P < 0.0001$. There were no significant differences in breathiness severity between speech samples categorized as control versus lesion(s).

The main effect of diagnostic label was also significant, but does not consider the accuracy of this label. Average breathiness severity was greatest when listeners rated speech samples with a diagnostic label indicating paralysis or paresis (M = 37.25; SEM = 1.43). There was, on average, a 3mm increase (95% CI = 0.78-4.5) in breathiness severity when a diagnostic label indicating paralysis/paresis was provided compared to the no diagnostic condition (M = 34.60; SEM = 1.25), P = 0.005. Breathiness was also increased, on average, by 4 mm (95% CI = 1.2-5.9) when this label was presented in comparison to diagnostic labels from the control group (M = 33.67; SEM = 1.44), P = 0.003. The interaction of the diagnostic label category and speaker category did not contribute significantly to the overall model for breathiness severity above and beyond that of speaker category, diagnostic labels, and dysphonia severity. In order to address the hypotheses related to the accuracy of the diagnostic label, selected post hoc comparisons were examined. First, significant differences between conditions in which listeners received no diagnostic information versus an accurate diagnosis of paralysis/paresis only approached significance, P = 0.09. When an inaccurate diagnosis of vocal fold paralysis or paresis was presented to listeners, speakers in the control group were perceived to be more

breathy compared to ratings obtained when an accurate diagnostic label was presented, $P = 0.02$. The largest difference in average perceived breathiness severity across all conditions, on average, was a 7 mm difference (CI = 3.1-11.4) when an accurate diagnosis of paralysis versus an inaccurate diagnostic label from the control group was presented to listeners, $P = 0.001$. These contrasts are presented in Figure 2.

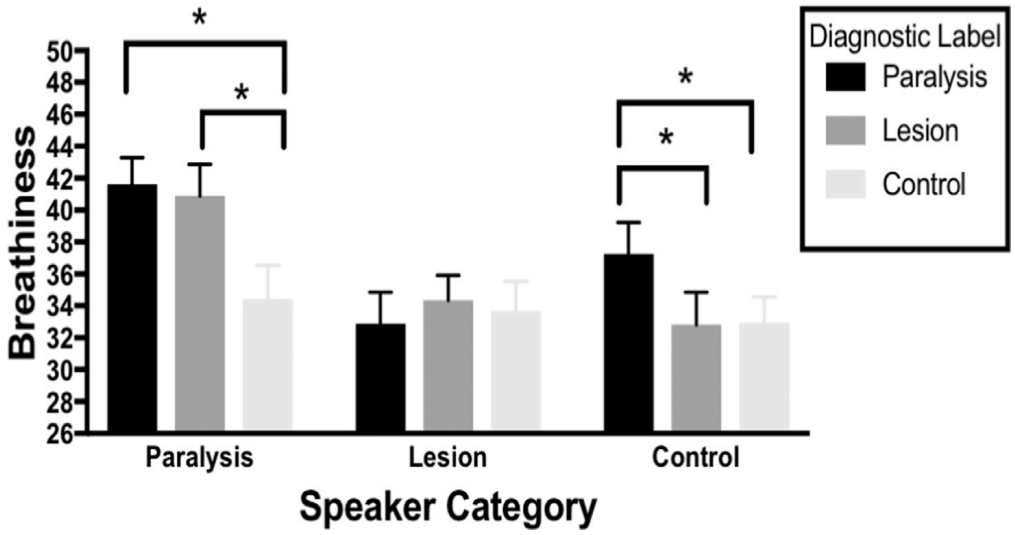


FIGURE 2. Breathiness severity by speaker category and diagnostic label.

DISCUSSION

Auditory-perceptual evaluation of voice quality is an important part of a comprehensive voice assessment. In a clinical setting, auditory-perceptual judgments often are made when information about the patient, including referring diagnosis, is known. Eadie et al.[10] found an effect of knowing an accurate medical diagnosis on perceived voice severity in novice and experienced listeners. The current study investigated the effects of both inaccurate and accurate medical diagnoses on perceptual judgments of breathiness and roughness in novice clinicians.

This is an important consideration, as evidence suggests that a referring medical diagnosis often may be inaccurate or nonspecific in a clinical setting.[12]

Based on previous findings,[10] it was hypothesized that only one expected dimension (breathiness or roughness) would increase in response to knowing specific diagnostic information when listeners made their judgments. That is, it was expected that listeners' perceptions of roughness would increase when they were provided a diagnostic label of vocal fold lesions, and that similar increases in breathiness would be observed when listeners made their judgments in the presence of specific diagnostic labels of vocal fold paralysis/paresis. Listeners were not expected to be susceptible to confirmation biases when speakers' diagnoses were within the control category; for speakers in this category, there were no expectations about whether there might be any changes in perceived breathiness or roughness. Yet, all of these hypotheses were generated on what was known when an accurate diagnosis was presented.[10] How inaccurate diagnoses might affect listeners' judgments was the particular focus of this investigation.

Overall, results from this study revealed that there was a greater effect of inaccurate diagnoses than accurate diagnoses on listeners' perceptions of roughness and breathiness. For example, the greatest significant differences in voice severity when an accurate diagnosis versus no diagnosis was presented was 4 mm (on a 100-mm VAS), whereas a 12-mm difference in severity was observed when speakers were rated with different inaccurate diagnoses (Table 1). Interestingly, when listeners were presented with diagnostic labels indicating either lesion(s) or paralysis/paresis (whether accurate or not), perceptions of roughness increased across speaker groups. However, consistent with the previous study,[10] increases in perceived breathiness were observed only in response to knowing that speakers had a diagnosis of vocal fold

paralysis/paresis (whether accurate or not). In contrast, when listeners were presented with inaccurate diagnostic labels from the control group, listeners' judgments of roughness and breathiness decreased in severity. Across all conditions (no diagnosis, accurate, and inaccurate diagnosis), judgments of roughness severity were more susceptible to bias compared to breathiness. How these results compare to previous literature is discussed next, along with considerations for clinical practice and future research.

Confirmation biases in evaluating dysphonia: effects of accuracy and voice dimension.

When only accurate diagnostic labels were considered, this study directly replicated findings from Eadie et al.[10] For example, listeners in this study significantly increased their perceptions of roughness (average = 4 mm), but not breathiness, when they made judgments in the presence of known accurate diagnoses of lesion(s) compared to no known diagnosis. Similar to the previous study, there also were trending increases in perceived breathiness (average = 2 mm), but not roughness, when judgments were made in conjunction with an accurate diagnosis of paralysis/paresis compared to no known diagnosis. These increases in listeners' perceptions of voice severity were similar to the 5mm (roughness) and 5mm (breathiness) increases observed in the no diagnosis to accurate diagnosis conditions in the previous study.[10] Findings for the control group were also consistent with previous findings; there were no significant differences in perceived breathiness or roughness when listeners made judgments in the presence of an accurate diagnosis for this group of speakers.[10] For example, knowing that a speaker had a true diagnosis of muscle tension dysphonia did not appear to create any confirmation biases for judgments of roughness or breathiness in the present group of listeners.

When both accurate and inaccurate diagnoses were considered in this study, several new findings were revealed. The preliminary hypothesis that voice severity would increase only along

specific voice dimensions was not maintained, specifically for the roughness dimension. Instead, speakers presented with diagnostic labels (either accurate or inaccurate) of vocal fold lesion(s) or vocal fold paralysis/ paresis were perceived to be significantly more rough compared to speakers presented with no diagnosis or a diagnosis from the control group. As expected, when listeners judged speakers with a true diagnosis of vocal fold paralysis in the presence of inaccurate labels of vocal fold lesions, their perceptions of roughness significantly increased. However, when listeners judged speakers with true diagnoses of vocal fold lesions in the presence of inaccurate labels of vocal fold paralysis/paresis, their perceptions of roughness also increased (Figure 1). Both of these results are consistent with an overpathologizing bias, which was not observed for speakers in the control group. In contrast, when speakers with true diagnoses of lesions or vocal fold paralysis/paresis were judged in the presence of inaccurate labels indicating a diagnosis in the control group (e.g., functional dysphonia), perceptions of roughness significantly decreased. These observations are consistent with a minimizing bias.

In contrast to the effects observed for ratings of roughness, listeners were less susceptible to biases from knowing a speaker's diagnosis when judging breathiness. Speakers who had a true diagnosis of vocal fold paralysis/paresis were perceived to be more breathy than speakers in the other groups, regardless of the accuracy or presence of a diagnostic label. The only condition that resulted in a confirmation (and overpathologizing) bias for breathiness occurred when speakers in the control group were judged in conjunction with an inaccurate diagnosis of vocal fold paralysis/paresis. That is, speakers in the control group were perceived as more severely breathy when listeners believed the speakers had a diagnosis of vocal fold paralysis/paresis. Consistent with the effect observed for roughness, speakers with true diagnoses of vocal fold

paralysis/paresis also were perceived as less significantly breathy when they were judged in the presence of inaccurate labels from the control group (e.g., a 7 mm difference).

Overall, results of this study should be interpreted in light of two main findings. First, this study revealed that inaccurate diagnoses had a stronger effect than accurate diagnoses on auditory-perceptual judgments of dysphonia in novice listeners. For example, differences as great as 12 mm and 7 mm for roughness and breathiness, respectively, were observed for speakers presented with different inaccurate diagnostic labels. In contrast, differences ranged from 2 mm to 4 mm when accurate diagnoses were known compared to unknown. Although the differences observed in the presence of inaccurate diagnoses (7-12 mm) are modest, one could argue that these differences, especially at the mild end of a 100-mm scale, could affect the validity of measuring outcomes that are assumed to be independent from knowing a medical diagnosis. Further research that establishes the clinical meaningfulness of these differences also needs to be performed.

Studies in other fields of medicine have revealed different effects of accurate and inaccurate referring diagnoses on other clinical assessment tools that also depend on visual and auditory-perceptual judgments for interpretation.[68] One main difference in these other studies is that they mostly include diagnostic accuracy as the primary outcome variable. For example, Hatala et al.[6] reported that the diagnostic accuracy of visual-perceptual judgments used in electrocardiogram interpretation increased by up to 12% when accurate versus no diagnostic information was presented. In contrast, diagnostic accuracy decreased by up to 25% when the inaccurate diagnosis was suggested in comparison to no diagnostic information. Hatala et al.[6] also found that the overall effect on diagnostic accuracy was greatest when an inaccurate versus accurate diagnosis was presented. Overall, these results are consistent with findings of this study,

which also showed a stronger effect of inaccurate versus accurate information on perceived voice quality. The results have important clinical implications, since patients with voice disorders may sometimes be evaluated in the presence of inaccurate referring diagnoses. Knowing how biases affect clinical judgments is the first step to controlling the potentially negative consequences of these effects in practice.[20,21]

The second main finding from this study relates to how diagnostic labels differentially affected judgments of roughness and breathiness. Specifically, judgments of roughness appeared to be more susceptible to biases than judgments of breathiness. Considering data from previous studies as well as the present study, we propose that listeners' vulnerability to bias in auditory-perceptual judgments of dysphonia is a function of both the strength of the expected bias and listeners' instabilities in making these types of judgments. For example, despite similar patterns of differences in perceived breathiness for accurate and inaccurate diagnoses, overall variability in perceived breathiness was reduced across conditions. This is one reason that the accuracy of diagnoses did not significantly contribute to the model for breathiness, but did contribute to the model for roughness. Additionally, measures of both intra- and inter-rater reliability and agreement were consistently higher for ratings of breathiness versus roughness in this study. This finding supports the contention that if a perceptual dimension is more reliable, it might be less susceptible to error. However, future research is needed to address these areas.

Finally, it is interesting to note that in this study, inaccurate labels that included diagnoses represented in the control group resulted in decreased perceived severity of roughness and breathiness in a number of conditions. These reductions in perceived severity of dysphonia were possibly related to the inclusion of the diagnostic label, "clear larynx," in the control diagnostic category. This diagnostic label might have led to expectations for more "normalized" voice

quality, possibly explaining the observed minimizing bias. These results have important implications, which will be explained next.

Future directions and limitations

Results of this study have several possible research and clinical implications, as well as limitations that need consideration. First, graduate student clinicians are frequently included in perceptual experiments evaluating voice quality.[1] In this study, they appeared to be susceptible to cognitive biases generated by presence of clinical information. This factor should be considered when designing future studies and interpreting results. In addition, while experienced raters also have been shown to be susceptible to biases,[6-8,10] these effects need further study in experienced clinicians.

Second, because it is known that inaccurate referring diagnoses may be somewhat common in clinical settings[12] and that knowledge of inaccurate diagnoses might affect auditory-perceptual judgments (in particular, judgments of roughness), blinding listeners to the referring diagnosis might help reduce the effect that a potentially inaccurate diagnosis could have on perceptual judgments. Since this is not always feasible in a clinical setting, routine voice recordings would allow comparison of ratings from clinicians blinded to diagnostic information, and those obtained when diagnostic information is known. The inclusion of objective (e.g., acoustic) measures to corroborate auditory-perceptual judgments might also support the validity of these measures, as acoustic measures are not subject to these types of biases.

Third, we must consider results from this study relative to the nature of the included speech samples. For example, studies have shown that speech samples that are mild to moderate in severity are more susceptible to bias.[10] This study design was limited in its interpretation of a three-way interaction among diagnostic label, speaker category, and severity. Therefore, how

dysphonia severity might have affected these results needs further investigation. As previously mentioned, we hypothesize that the instability of these types of auditory-perceptual judgments may increase listeners' susceptibility to biases in performing these types of judgments. How biases might affect judgments of voice qualities that have been shown to be less reliable, such as vocal strain,[22] also need further study. Any approach that purports to increase rater reliability may help stabilize these outcomes and render them less sensitive to these effects.[19,23]

Finally, future research should determine how confirmation biases might affect other types of perceptual judgments, such as visual-perceptual judgments of videolaryngostroboscopic parameters, and whether they affect diagnostic or treatment recommendations. While Teitler[9] examined a related question, the results are limited to this one study and warrant future study. All of these areas have meaningful implications for clinical practice, as both visual- and auditory-perceptual outcomes are highly valued measures and commonly used in clinical settings.

CONCLUSIONS

This study revealed that knowledge of inaccurate diagnostic information prior to making auditory-perceptual judgments of dysphonia had an overall stronger effect than knowing accurate diagnoses among a group of novice listeners. Findings also suggested that both accurate and inaccurate referring diagnoses had a stronger influence on roughness versus breathiness severity ratings. Results from this study have implications for both research and clinical practice. This topic warrants future study in how this type of information might also affect experienced clinicians' perceptual judgments across a variety of domains during voice assessment.

References

1. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. J Speech Hear Res. 1993;36:21.
2. Kahneman D, Kahneman D, Slovic P, Tversky A. Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press; 1982.
3. Darley M, Gross P. A hypothesis-confirming bias in labeling effect. J Pers Soc Psychol. 1983;44:20–33.
4. Langer EJ, Abelson RP. A patient by any other name clinician group difference in labeling bias. J Consult Clin Psychol. 1974;42:4–9.
5. Lopez SR. Patient variable biases in clinical judgment: conceptual overview and methodological considerations. Psychol Bull.1989;106:184–203.
6. Hatala R, Norman GR, Brooks LR. Impact of a clinical scenario on accuracy of electrocardiogram interpretation. J Gen Intern Med. 1999;14:126–129.
7. Leblanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. Acad Med J Assoc Am Med Coll. 2002;77(10 Suppl):S67– S69.
8. Shikino K, Ikusaka M, Ohira Y, et al. Influence of predicting the diagnosis from history on the accuracy of physical examination. Adv Med Educ Pract. 2015;6:143–148.
9. Teitler N. Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. J Voice. 1995;9:95–105.
10. Eadie T, Sroka A, Wright DR, Merati A. "Does knowledge of medical diagnosis bias auditory-perceptual judgments of dysphonia? J Voice. 2011;25:420.
11. Ramig LA. Effects of examiner expectancy on speech ratings of individuals with cleft lip and/or palate. Cleft Palate J. 1982;19:270–274.
12. Cohen SM, Kim J, Roy N, Wilk A, Thomas S, Courey M. Change in diagnosis and treatment following specialty voice evaluation: a national database analysis. Laryngoscope. 2015;125:1660–1666.
13. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. Am J Speech Lang Pathol. 2009;18:124–132.
14. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. J Voice. 2007;21:576–590.
15. Fairbanks G. Voice and Articulation Drillbook. 2nd ed. New York, NY: Harper & Row; 1960.
16. IBM, IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.
17. Shek DTL, Ma CMS. Longitudinal data analyses using linear mixed models in SPSS: concepts, procedures and illustrations. Sci World J. 2011;11:42–76.
18. Janke SJ, Tinsley FC. Introduction to Linear Models and Statistical Inference. Hoboken, NJ: John Wiley & Sons, Inc; 2005.
19. Chan KMK, Yiu EM-L. The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res. 2002;45:111–126.

20. Bornstein BH, Bornstein BH, Emler AC, Chapman GB. Rationality in medical treatment decisions: is there a sunk-cost effect? Soc Sci Med. 1999;49:215–222.
21. Gruppen L, Gruppen LD, Margolin J, Wisdom K, Grum CM. Outcome bias and cognitive dissonance in evaluating treatment decisions. Acad Med. 1994;69:S57–S59.
22. Webb AL, Carding PN, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. Eur Arch Otorhinolaryngol. 2004;261:429–434.
23. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. J Voice. 2006;20: 527–544.

Chapter 3

Does the Accuracy of Case History Affect Interpretation of Videolaryngostroboscopic Exams?

Abstract

Objective: To determine the effect of initial diagnostic hypotheses on clinicians' 1) detection and perceived severity of abnormalities, and 2) clinical impressions and treatment recommendations for individuals with and without voice disorders following interpretation of videolaryngostroboscopy (VLS).

Methods: Thirty-two experienced speech-language pathologists and otolaryngologists specializing in voice disorders read case histories prior to interpreting exams. Case histories suggested specific accurate or inaccurate laryngeal diagnoses, or a control scenario that suggested a normal larynx. The effects of the accuracy of case histories on perceived severity of associated visual-perceptual parameters, clinical impressions, and treatment recommendations were examined.

Results: Significant increases in perceived severity of posterior laryngeal appearance ($P < 0.05$) and mucosal wave ($P < 0.02$) were observed when these abnormalities were suggested by case histories. Overall agreement with clinical impressions improved from 49% to 72% when the case history was consistent with the examination. Case histories (accurate and inaccurate) indicating voice symptoms predicted recommendations for treatment above and beyond that of VLS presentation alone, $P < 0.001$.

Conclusion: Case histories suggesting specific abnormalities significantly affected severity ratings for two of three associated visual-perceptual parameters selected as primary outcome measures. Accurate case histories suggesting specific abnormalities increased the probability of detection and perceived severity. Inaccurate case histories led to false-positive findings and failures to detect abnormalities or to interpret them as less severe. Case histories affected visual-perceptual judgments and contributed to decisions about clinical impressions and treatment.

INTRODUCTION

Videolaryngostroboscopy (VLS) is a highly valued tool that is important for the assessment, diagnosis, and treatment of voice disorders.[1–3] During a typical clinical examination, an endoscope paired with a continuous light source is primarily used to obtain specific measures of appearance and gross movement, whereas the stroboscopic effect is used to address a variety of vibratory characteristics. Although it is widely acknowledged that specific parameters are most validly measured with or without stroboscopy, the abbreviation VLS is commonly used to refer to the examination as a whole[4]; this terminology will be used as such throughout this article. Despite the tool's importance to otolaryngologists (ENTs) and speech-language pathologists (SLPs), commonly rated VLS parameters are subjective and might be affected by rater expertise, rating tools, and severity of dysphonia, among other factors.[4–6]

Although some sources of variability in these visual-perceptual ratings are related to random error, others might be more systematic. A systematic deviation from the norm in perceptual judgment or reasoning is known as a cognitive bias.[7] One clinically significant source of cognitive bias is confirmation bias, which occurs when information about a patient predisposes a clinician to expect, and consequently seek out or interpret, available information to confirm an initial diagnostic hypothesis.[8] An initial diagnostic hypothesis can be affected by the accuracy of a referring diagnosis[9] and/or case history. Some confirmation biases are advantageous but in other cases may lead to failure to detect or correctly interpret findings, resulting in misdiagnosis and/or poorer quality treatment recommendations.[10–12] Thus, a better understanding of these issues is meaningful to clinicians who use VLS to direct patient care.

The effect of fictional case history information on clinicians' ratings of VLS parameters has been examined in one prior study.[13] Case histories that were negative (e.g., patient was a

smoker, dehydrated, voice misuse) or positive (e.g., patient was hydrated, adhered to treatment) were presented in conjunction with videos. Results revealed that the perceived abnormality of 12 of 13 VLS parameters was increased in severity when clinicians were presented with negative compared to positive histories. The absence of case history also resulted in increased perceived severity of many parameters in comparison to these fictional case histories, but the study design limited additional post hoc comparisons of this control group. The results demonstrated that clinicians were susceptible to confirmation bias for some visual-perceptual measures of vocal function.

Although the prior study provided evidence that fictional clinical information affected the perceived severity of VLS parameters, it did not examine the effect of the accuracy or consistency of the case history with the VLS exam. It also did not investigate whether differences in perceived severity of these visual-perceptual measures might influence clinical impressions, diagnosis, or treatment recommendations. The primary aim of the present study was to determine the effect of specific accurate and inaccurate diagnostic hypotheses on experienced clinicians' detection and interpretation of VLS abnormalities. The second aim was to determine whether differences in visual-perceptual ratings might affect clinical impressions and treatment recommendations in those with and without voice disorders.

MATERIALS AND METHODS

The institutional review board at the University of Washington, Seattle, Washington, approved this study.

Case History Development and Stimuli Selection

Case history information suggestive of three diagnostic hypotheses and one control scenario was generated. A consensus panel consisting of four experienced SLPs (7–23 years experience) independently provided confidence ratings[3] to ensure that case histories were associated with expectations for specific abnormal visual-perceptual features and that the control history was not associated with any abnormal features. The selected case histories were suggestive of left vocal fold (LVF) paresis, laryngeal pharyngeal reflux (LPR), hemorrhage/hemorrhagic lesion(s), and normal larynx (control). These diagnoses were selected because each condition is associated with at least one abnormal feature that is perceptually distinct from the other diagnoses; the design thereby facilitated a clear interpretation of the effect of case history independent of other effects.

The consensus panel then selected VLS exams that were consistent with each case history from among those obtained during standard clinical evaluations at the University of Washington Speech and Hearing Clinic. Panel members independently rated the severity of all VLS parameters for each video. One unique perceptual feature present in each experimental video and associated with expectations for abnormal findings generated by each case history was identified as the primary outcome measure (LVF paresis: LVF mobility, LPR: posterior laryngeal appearance, resolving hemorrhagic lesion/hemorrhage: right mucosal wave). The consensus panel confirmed that abnormal parameters selected as primary outcome measures were unique to each experimental video such that the perceived abnormality of that parameter in any other video could be attributed to the history alone. All primary outcome measures were rated as normal for the control video (normal larynx). Experimental videos represented pathology that was mild to moderately-severe.

All VLS exams were recorded using a 70-degree Kay Pentax (Montvale, NJ) rigid laryngoscope, a 9200C processor, 9100B light source, and Toshiba (Irvine, CA) 3CCD camera. A single instrumentation method (rigid exam) was used to control the potential confounding effect of instrumentation on visual-perceptual parameters while ensuring high-quality videos adequate for rating purposes.[6,14] Recordings were captured at the standard 30 frames per second; auditory information was retained. Videos included three tokens of sustained /i/ phonation at comfortable pitch and loudness levels, ascending and descending pitch glides, high-pitched phonation, and a laryngeal diadochokinetic task. Videos were approximately 2 minutes in duration and included an adequate number (i.e., >3) of videostroboscopic glottal cycles.[15] Videos were saved as .avi files with limited compression and transferred to a laptop computer with a 13.3 inch display.

Participants

Nine ENTs and 23 SLPs (N = 32; average clinical experience = 15.75 years; standard deviation = 8 years; range = 3–38 years) with at least 3 years of specialized clinical experience evaluating and treating voice disorders participated. These clinicians were recruited at local, regional, and national conferences and through professional contacts. To facilitate recruitment efforts, VLS exams were viewed on a laptop in a location and at a time that was convenient for individual participants. Participants were quasi-randomized into four rater groups (N = 8) so that at least two ENTs were included in each group to control for professional background. Based on a prior study[13] the sample size and power were deemed adequate for measuring any observable effects. Participants completed questionnaires about professional background, years of experience, practice setting, geographic location, and use of rating tools in clinical practice. Across four rater groups, there were no significant differences in the proportion of ENTs to SLPs

($P = 0.927$) or clinical experience ($P = 0.470$). Practice settings included outpatient clinics (6.3%), private practice (18.7%), hospital-based outpatient setting (37.5%), and academic practice (37.5%). Most participants reported using no specific VLS rating tool (72%) as part of routine clinical practice. Others reported using a facility-specific rating form (19%) or a previously published rating form (9%).

Rating Procedure

Case histories that suggested either accurate or one of three inaccurate diagnostic hypotheses were presented to participants before viewing four VLS exams to determine the effect of case history information on VLS interpretation, diagnostic impressions, and treatment recommendations. In an effort to control the effect of the rating task on outcomes, raters received a detailed rating form that included operational definitions and visual examples of many visual-perceptual parameters (Appendix B), similar to the Voice-Vibratory Assessment With Laryngeal Imaging tool.[16] They were permitted to refer to this rating form as needed. Glottal closure pattern was rated using a nominal scale, and additional parameters were rated using 4-point ordinal scales (0 = normal, 1 = mild, 2 = moderate, 3 = severe). The VLS rating form included nine parameters to ensure inclusion of all primary outcome measures, reduce redundancy in rating parameters, and maintain blinding of raters. Prior to the rating phase, participants were presented a 1-minute familiarization video. This novel video included VLS examinations that were similar to the types and ranges of severity of the stimuli included in the rating tasks. Each group of raters was presented with a different case history before examining each video (see Table I).

TABLE I.
Case Histories Presented With Each Video for Each Rater Group.

	Rater Groups			
	Group 1	Group 2	Group 3	Group 4
Video 1	**History A**	History C	History D	History B
Video 2	History D	**History B**	History A	History C
Video 3	History B	History D	**History C**	History A
Video 4	History C	History A	History B	**History D**

Bold depicts accurate video clinical history combinations for each rater group.

TABLE I Case Histories Presented With Each Video for Each Rater Group

Raters within each group were presented the videos in one of four possible presentation orders to control for order effects.

Individual raters viewed the videos a maximum of two times, were instructed to pause and resume the video as needed, and were provided unlimited time to complete all ratings. Auditory information was retained in the video and presented via headphones (Samson RH600, Samson Technologies, Hicksville, NY) at a comfortable listening level. After rating all parameters, raters provided diagnostic impressions and treatment recommendations along with confidence ratings. Upon completion, raters were debriefed about the study's specific purpose. Sessions lasted approximately 30 minutes.

RESULTS

Rater Reliability

Intra-rater reliability was assessed by presenting the accurate case history and video combination to raters a second time. Intra-rater reliability was calculated using Spearman's rank correlation coefficient and percentage (%) exact agreement of severity rating (see Table II). Measures of interrater reliability were calculated using an extension of Fleiss kappa[17]—a "chance-corrected" measure of agreement for multiple raters for ordinal weighted linear data (R studio version 1.1.44)—and were compared to Landis and Koch's[18] standards for rater reliability. Exact agreement with mode for all parameters across all videos X history combinations was also calculated (see Table III). Because glottal closure was measured using nominal data, interrater reliability was measured using Fleiss Kappa k (unweighted) = 0.573 (95% confidence interval = 0.229–0.885) and percent absolute agreement with mode = 74.25%. There were no statistically significant differences in interrater reliability for the primary outcome measures for raters with accurate versus inaccurate histories(ps >0.05).

TABLE II.
Intra-rater Reliability for 50 Percent of VLS Dimensions, Including Primary Outcomes Measures.

	Spearman's Rs	% Exact Agreement
Posterior laryngeal appearance	0.748	0.813
Vocal fold mobility: right	0.999	0.969
Vocal fold mobility left	0.923	0.969
Mucosal wave: right	0.758	0.813
Mucosal wave: left	0.767	0.688
Erythema: right	0.949	0.969
Erythema: left	0.946	0.906
Glottal gap size	0.863	0.844

VLS = videolaryngostroboscopy.

TABLE III.
Interrater Reliability and Percentage Overall Exact Agreement With Mode.

VLS Rating Dimension	Weighted Kappa k^w	95% Confidence Interval	% Overall Exact Agreement with Mode
Posterior laryngeal appearance	0.179	0.079–0.231	49.23%
Mobility right	0.830†††	0.661–1.000	92.20%
Mobility left	0.727††	0.399–0.926	79.70%
Erythema right	0.626††	0.503–0.748	81.25%
Erythema left	0.690††	0.587–0.808	78.20%
Lesion size right	0.754††	0.619–0.880	89.85%
Lesion size left	0.597†	0.593–0.914	82.05%
Medial edge right	0.644††	0.547–0.721	79.73%
Medial edge left	0.597†	0.438–0.718	67.20%
Mucosal wave right	0.685††	0.449–0.885	73.90%
Mucosal wave left	0.611††	0.356–0.866	71.13%
Glottal gap size	0.522	0.233–0.812	75.80%
Phase symmetry	0.498	0.298–0.726	71.88%

VLS = videolaryngostroboscopy. †moderate reliability (.55–.60)
††substantial reliability (.61–.80) †††almost perfect reliability (>.81)

Descriptive Statistics

Video Stimuli

All participants' ratings for the primary outcome measures (LVF mobility, posterior laryngeal appearance, right mucosal wave) for each video are presented in Figure 1. The ratings reflect the number of times each severity category was selected (i.e., a frequency measure) across all video X history combinations for comparison with consensus panel judgments. For example, the control video was judged by the consensus panel to be normal for all VLS outcomes. Raters who judged any parameters as abnormal for this video disagreed with the consensus panel. Likewise, each experimental video was associated with only one abnormal outcome measure so that frequencies >0 in other parameters reflect false positives, and frequencies <32 for the associated parameter indicate that raters failed to detect the abnormality. Across all video X history combinations, the severity of LVF mobility and right mucosal wave was significantly increased in the associated experimental videos compared to all other videos, $P < 0.001$. The severity of posterior laryngeal appearance was significantly increased in the associated experimental video compared to video 1 (control) and video 2 (LVF mobility), $P < 0.02$, but was not significantly different from video 4 (right mucosal wave). Ordinal regression models used to examine the effect of history on these primary outcome measures are presented below.

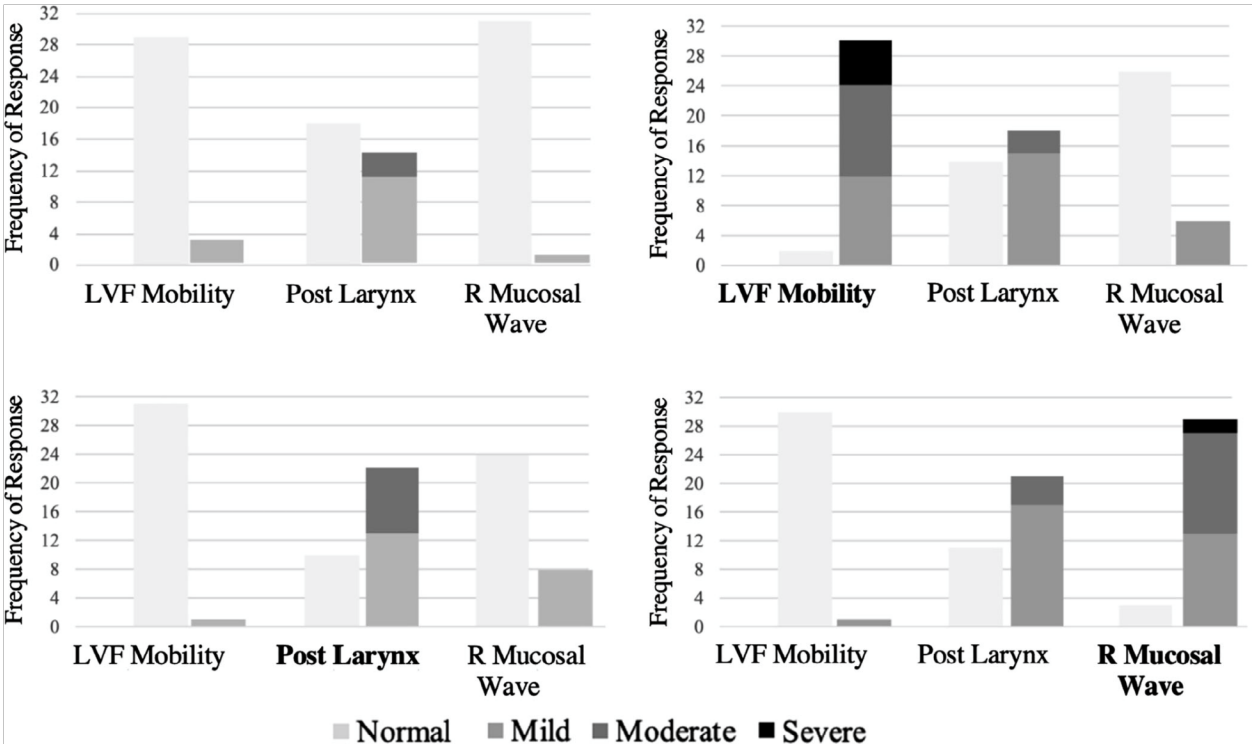


Figure. 1. Frequency of rating severity for all primary outcomes across all video X history combinations. Ratings that are normal are in left-hand bars for each parameter, and ratings that are non-normal (mild, moderate, severe) are grouped in one right-hand bar for each parameter. Associated abnormal parameter for each video is bolded. Top left panel: video 1 (control); upper right panel: video 2 (left vocal fold mobility); lower left panel: video 3 (posterior laryngeal appearance); lower right panel: video 4 (right mucosal wave).

Effect of History on Primary Outcome Measures

SPSS (version 24, IBM Corp., Armonk, NY) for Mac[19] was used to create separate ordinal regression models to determine whether case histories predicted the severity of each primary outcome variable (LVF mobility, posterior laryngeal appearance, and right mucosal wave). Preliminary models for all outcome variables included video, case history, and all video X case history combinations. Interactions were removed from models when there were too many limitations in response categories. For all models, the chi-square statistic was significant ($P < 0.001$), indicating that the final models gave a significant improvement over the baseline were as

follows: LVF mobility, pseudo $r^2 = 0.662$; posterior laryngeal appearance, pseudo $r^2 = 0.217$; right mucosal wave, pseudo $r^2 = 0.615$.

To address the first aim, ordinal regression models were used to test whether case history information influenced the perceived severity (normal, mild, moderate, severe) of parameters across all videos (see Fig. 2A). Results showed that when a history suggested LVF paresis (history B), the increased severity of LVF mobility abnormality was not statistically different from when any other history was presented. In contrast, the effect of history on severity ratings of posterior laryngeal appearance (history C) was found to significantly differ across video X history combinations. Results show that posterior laryngeal appearance was consistently rated as more severely abnormal when a diagnostic hypothesis suggested LPR compared to the control history for all four videos, $P < 0.05$. For the associated experimental video, the severity of posterior laryngeal abnormality was greatest when an accurate case history suggested LPR versus any other inaccurate history, $P < 0.05$. Lastly, when a history suggested hemorrhage/hemorrhagic lesion, right mucosal wave significantly increased in severity compared to all histories, $P < 0.02$, except the control history. Results are presented in Figure 2A, which shows the probability of detecting (true and false positives) and interpreting the severity of the primary outcome measures by case history. Figure 2B shows the probability of severity ratings for the associated experimental and control videos by case history, respectively.

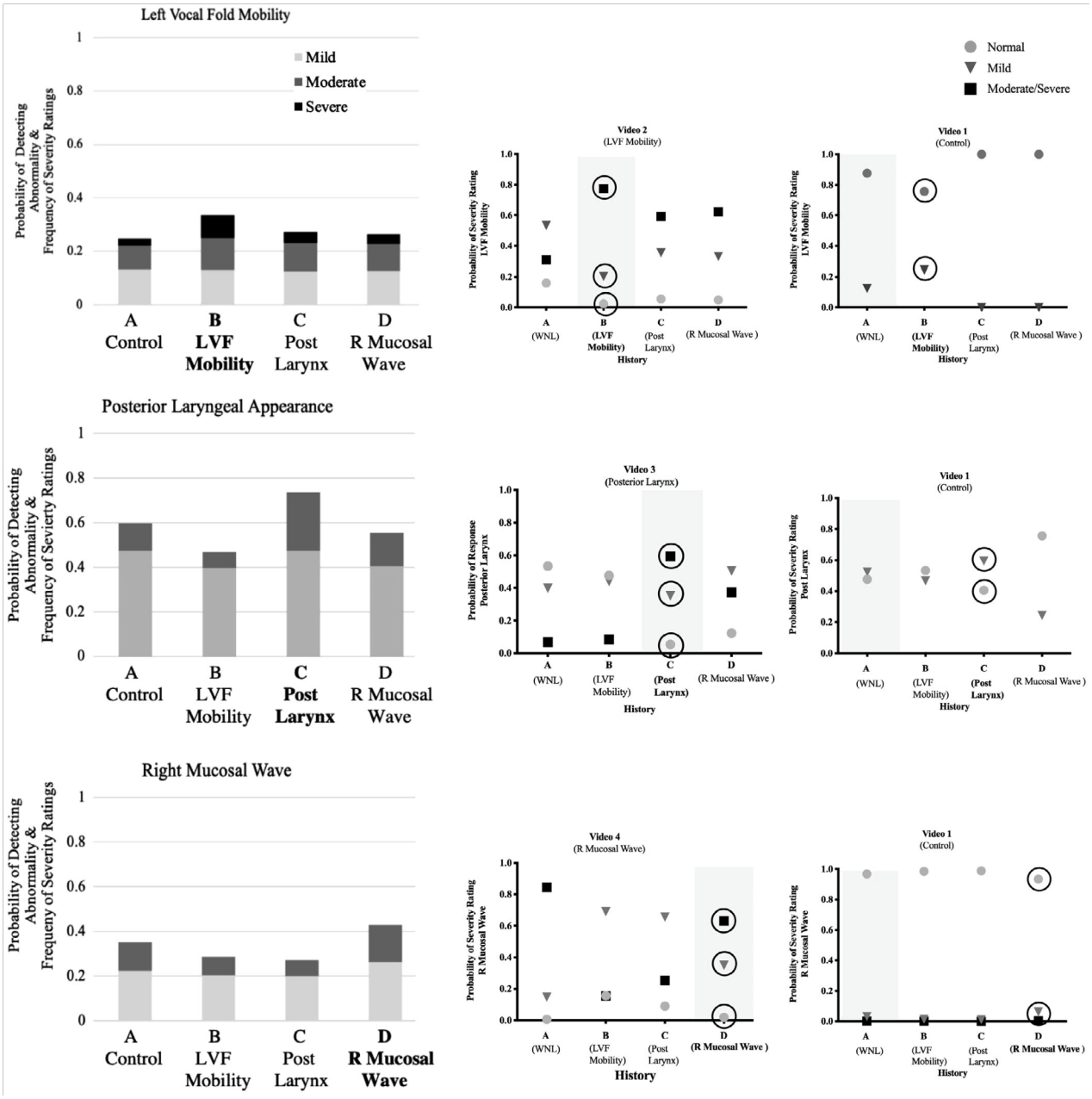


Figure 2. (A: left panels A, B, C) Probability of detecting abnormality (true and false positives) and frequency of severity ratings for primary outcome measures as a function of the case history across videos. Case histories that align with the expected abnormal parameter are bolded along the x-axis.

Fig 2. (B: right panels A, B, C) Results for the experimental video and control video results are shown for comparison purposes. The accurate case histories are shaded. History associated with the abnormality is bolded along the x-axis and circled.

Effect of History on Diagnostic Impressions and Treatment Recommendations

The second aim examined the effect of case history on "accuracy" of diagnosis and treatment recommendations. Results revealed a moderately significant relationship between history and diagnosis (Cramer's $V(128) = 0.373$, $P < 0.001$). Overall agreement with the consensus panel's clinical impressions for all video X history combinations was 56%. When histories were inconsistent with the video presentation, exact agreement was 49%; in contrast, when histories were presented with the associated experimental video, agreement increased to 72%. Overall, 88.5% of participants rated their confidence in clinical impressions as confident or really confident. There were no significant differences in confidence when case history was accurate versus inaccurate ($P = 0.08$). Clinical impressions provided by all raters are presented in Figure 3.

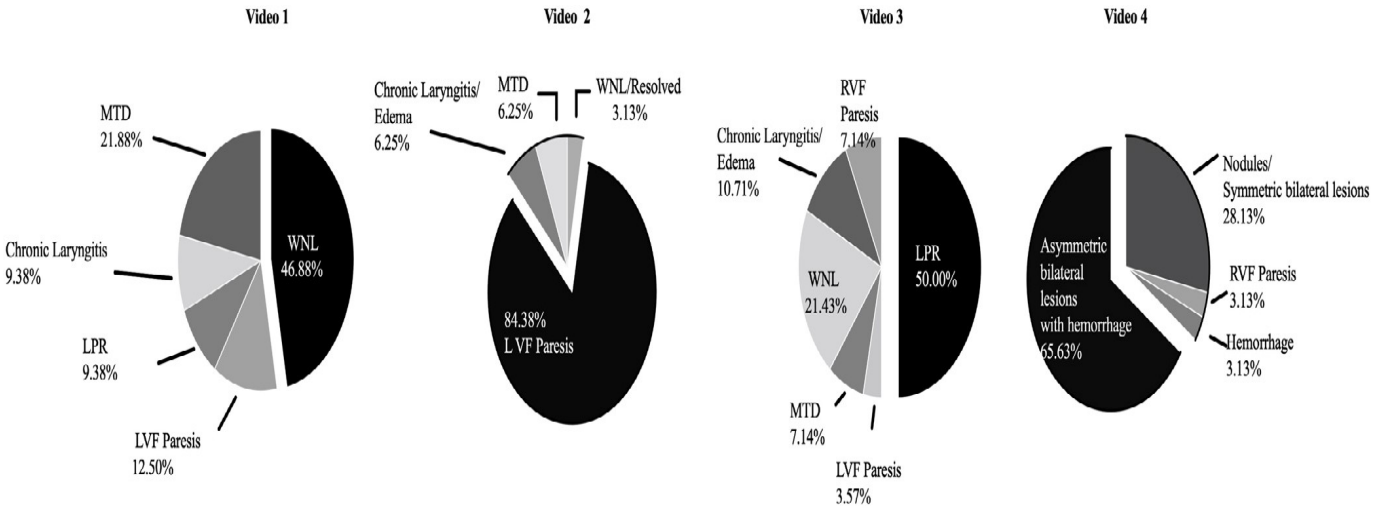


Figure 3. Diagnostic impressions for all video X history combinations. Black sector depicts agreement with consensus panel selection.

Treatment recommendations were categorized as 1) no treatment, surveillance, monitor (28%); 2) voice therapy (52%); 3) pharmaceutical/behavioral management of LPR (9%); 4) surgery/injection (2%); 5) steroids (9%) or; 6) further assessment (<1%). Overall, 96.1% of participants indicated that they were confident or really confident with treatment recommendations, irrespective of whether the history matched the video ($P = 0.50$).

Two multiple logistical regression models with sequential predictor entry were used to determine the contribution of case history above and beyond that of the video stimuli alone to predict 1) voice therapy recommendations; or 2) recommendation for surveillance, monitoring, or no treatment. The overall hit rate for predicting voice therapy recommendations significantly improved from 68% to 74.2% when history was added to the model with the video, $\chi^2 = 31.64$, $P < 0.001$. Accurate and inaccurate case histories suggesting voice symptoms were more likely to receive a recommendation for voice therapy compared to the control history that included no voice complaints. For predicting no treatment, the addition of history significantly increased the overall hit rate from 70.3% to 78.1% compared to the video alone, $\chi^2 = 32.24$, $P < 0.001$. Recommendations for no treatment were more probable when a case history suggested no voice symptoms (control), even when laryngeal abnormalities were observed.

DISCUSSION

Case history information has been shown to affect visual-perceptual judgments and decisions about diagnosis and treatment in many fields of medicine.[10,11] The focus of the present study was whether the accuracy of this information might affect visual-perceptual judgments of laryngeal function and structure, clinical impressions of diagnosis, and treatment recommendations in voice disorders.

Videolaryngostroboscopy Severity Ratings

In this study, clinicians were more likely to perceive the primary outcome measures as more severe when the preliminary case history suggested abnormality, consistent with one previous study.[13] Clinicians perceived posterior laryngeal appearance and mucosal wave as significantly more abnormal when the presenting case history supported these expectations. One explanation for this finding is that examiners were more likely to identify and/or interpret information that supported a preliminary diagnostic hypothesis and/or failed to observe and/or interpret findings that did not support this hypothesis, consistent with a confirmation bias.[20]

Importantly, the accuracy of a clinician's preliminary diagnostic hypothesis had both positive and negative effects. Overall, clinicians were more likely to detect the presence of abnormal features when they were provided an accurate case history that suggested this abnormality (see Fig. 2A). The impact of an inaccurate history also influenced raters in different ways. This is an important scenario to study because case history information may sometimes lead an examiner to develop plausible, although incorrect, diagnostic hypotheses that might affect visual-perceptual judgments. In this study, results showed that normal features sometimes were interpreted as abnormal. Some participants who were presented an inaccurate history suggesting LVF paresis erroneously identified abnormal LVF mobility in the control video. Inaccurate histories also resulted in a failure to detect abnormalities (see Fig. 2B). In other medical fields, a failure to detect abnormalities has been associated with visual search strategies, memory, and attention, suggesting that further study in voice disorders may be important.[21]

Some findings might also relate to specific visual-perceptual parameters. For example, although a history suggesting LPR influenced ratings of posterior laryngeal appearance,

clinicians were somewhat unreliable in performing these ratings, consistent with previous studies.[14,22] Abnormal posterior laryngeal appearance may be found in asymptomatic adults, indicating that information from VLS alone is not always sufficient to distinguish non-symptomatic and symptomatic individuals.[23,24] Further, although all videos were recorded using a rigid endoscope in this study, the endoscopic technique used to examine specific visual-perceptual features (e.g., gross vocal fold movement) might be another source of variance[14] that warrants consideration when generalizing these findings to clinical settings. Finally, it is also possible that case history information might help clinicians interpret the relevance of these abnormalities.[25] Thus, future studies examining how the presence/absence of information, such as case history and auditory information, might affect different visual-perceptual parameters appear warranted.

Clinical Impressions and Treatment Recommendations

Case history information also affected clinical impressions and treatment recommendations in this study. Using direct laryngoscopy as a gold standard, VLS information previously has been shown to increase examiner confidence and diagnostic accuracy of vocal fold lesions.[3] However, for many voice disorders, a gold standard for diagnostic accuracy does not exist. Instead, a high level of agreement among clinicians may be used as a surrogate standard.[26] In this study, clinicians' exact agreement with consensus panel clinical impressions was 49% when histories were inconsistent with the presented video, and it rose to 72% when histories were consistent. These results support the contention that accuracy of case history impacts VLS judgments and clinical impressions. Interestingly, rater confidence was unaffected

by the accuracy of case history information, suggesting that clinicians might have been unaware of this effect.

Although the overall results were consistent with this pattern, one exception involved the video with the most extensive laryngeal pathology (video 4). For this video, the high level of agreement with clinical impressions and increased probability of detecting mucosal wave abnormalities were similar for raters, who were provided an accurate history or the control history. Reduced plausibility of this video and control history combination might explain the different effect of the control history on this particular video. Some disagreements in clinical impressions were directly related to visual-perceptual "errors"; however, factors such as which specific laryngeal features are consistent with laryngeal diagnoses, as well as the relative importance assigned to case history and/or VLS information in determining the nature of voice symptoms, deserve further consideration. It is also important to note that some of these decisions are more controversial or complex than others, possibly rendering some diagnoses more susceptible to these effects. Future studies that include a larger number of diagnoses will help answer these questions.

Many factors contribute to decisions about treatment recommendations.[27] In this study, brief case histories that suggested voice symptoms have predicted recommendations for voice therapy above and beyond the video stimuli alone ($P < 0.05$). There was increased probability of recommending no treatment or surveillance when the control history suggested no voice symptoms, irrespective of abnormal findings. Results from this study are consistent with clinical practice. In the absence of malignancy, many clinicians do not recommend intervention to asymptomatic patients, even when laryngeal pathology exists.[28]

There are several clinical implications of this study. Strategies to reduce visual-perceptual errors and negative consequences associated with retaining inaccurate diagnostic hypotheses should be considered. Such strategies include using a checklist or standardized rating form or other methods to optimize search strategies, memory, and attention.[29,30] Surprisingly, few participants reported using standardized rating forms in their clinical settings. We postulate that the use of a standardized rating form and operational definitions, inclusion of experienced clinicians, unconstrained time, and the use of high-quality VLS exams might have mitigated the effect of clinical information on visual-perceptual "errors" in this study. Without these controls, it is likely that these types of visual-perceptual errors might be more pronounced in clinical settings.[31,32]

How results from this study might relate to clinical experience, profession, or practice setting also deserves consideration. For example, whereas ENTs and SLPs both use VLS, the purpose of their examinations in clinical settings might differ. Given that the results from this study showed an effect of case history on visual-perceptual ratings and clinical judgments as a group, additional studies that focus on specific rater characteristics appear warranted.

The most important implication of this study is that because case history affected the perceived severity of some VLS parameters, ratings should be performed blinded when used to measure treatment outcomes. Yet, rater blinding was reported in only 31% of studies using VLS as a primary treatment outcome measure.[33] However, it is difficult to reconcile the advantages of an accurate preliminary diagnostic hypothesis in detecting VLS abnormalities with the disadvantages of an inaccurate hypothesis during interpretation of an exam in a clinical setting when the purpose of such an exam is to assess laryngeal function and structure, render a diagnosis, and make treatment recommendations. Thus, it is important that clinicians and

researchers are aware of how cognitive biases affect visual-perceptual judgments because this may affect how VLS results are interpreted and how future investigations should be designed.

CONCLUSION

Provision of case histories suggesting different diagnostic hypotheses affected severity ratings, resulting in statistically significant differences for two of three parameters selected as primary outcome measures in this study. Accurate case histories suggesting specific abnormalities increased the probability of detecting these abnormalities and rating them as more severe. Inaccurate histories led to some false positives and failures to detect abnormalities. The moderate association between case history and clinical diagnostic impressions as well as treatment recommendations indicates that history is an important determinant of some clinical decisions in voice disorders.

References

1. Patel R, Awan S, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American Speech-Language Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. Am J Speech Lang Pathol 2018;27: 887–905.
2. Fritz MA, Persky MJ, Fang Y, et al. The accuracy of the laryngopharyngeal reflux diagnosis: utility of the stroboscopic exam. Otolaryngol Head Neck Surg 2016;155:629–634.
3. Paul BC, Chen S, Sridharan S, Fang Y, Amin MR, Branski RC. Diagnostic accuracy of history, laryngoscopy, and stroboscopy. Laryngoscope 2013; 123:215–219.
4. Bonilha HS, Desjardins M, Garand KL, Martin-Harris B. Parameters and scales used to assess and report findings from stroboscopy: a systematic review. J Voice 2018;32:734–755.
5. Rosen CA. Stroboscopy as a research instrument: development of a perceptual evaluation tool. Laryngoscope 2005;115:423–428.
6. Hosbach-Cannon CJ, Lowell SY, Kelley RT, Colton RH. A Preliminary quantitative comparison of vibratory amplitude using rigid and flexible stroboscopic assessment. J Voice 2016;30:485–492.
7. Itri JN, Patel SH. Heuristics and cognitive error in medical imaging. AJR Am J Roentgenol 2018;210:1097–1105.
8. Gurmankin AD, Baron J, Hershey JC, Ubel PA. The role of physicians. Recommendations in medical treatment decisions. Med Decis Making 2002; 22:262–271.
9. Cohen SM, Kim J, Roy N, Wilk A, Thomas S, Courey M. Change in diagnosis and treatment following specialty voice evaluation: a national database analysis. Laryngoscope 2015;125:1660–1666.
10. Leblanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. Acad Med 2002;77(suppl):S67–S69.
11. Hatala R, Norman GR, Brooks LR. Impact of a clinical scenario on accuracy of electrocardiogram interpretation. J Gen Intern Med 1999;14:126–129.
12. Bytzer P. Information bias in endoscopic assessment. Am J Gastroenterol 2007;102:1585–1587.
13. Teitler N. Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. J Voice 1995;9:95–105.
14. Milstein CF, Charbel S, Hicks DM, Abelson TI, Richter JE, Vaezi MF. Prevalence of laryngeal irritation signs associated with reflux in asymptomatic volunteers: impact of endoscopic technique (rigid vs. flexible laryngoscope). Laryngoscope 2005;115:2256–2261.
15. Hertegard, S. What have we learned about laryngeal physiology from highspeed digital videoendoscopy? Curr Opin Otolaryngol Head Neck Surg 2005;13:152–156.
16. Poburka BJ, Patel RR, Bless DM. Voice-vibratory assessment with laryngeal imaging (VALI) form: reliability of rating stroboscopy and high-speed videoendoscopy. J Voice 2017;31:513.e1–513.e14.
17. Mielke PW, Berry KJ, Johnston JE. Unweighted and weighted kappa as measures of agreement for multiple judges. Int J Manag 2009;26:213–223.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.

19. IBM Corp. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp; 2016
20. Jonas E, Schulz-Hardt S, Frey D, Thelen N. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. J Pers Soc Psychol 2001;80:557–571.
21. Kim YW, Mansfield LT. Fool me twice: delayed diagnosis in radiology with emphasis on perpetuated errors. AJR Am J Roentgenol 2014:202:465–470.
22. Branski R, Bhattacharyya N, Shapiro J. The reliability of the assessment of endoscopic laryngeal findings associated with laryngopharyngeal reflux disease. Laryngoscope 2002;112:1019–1024.
23. Elias ME, Sataloff RT, Rosen DC, Heuer RJ, Spiegel JR. Normal strobovideolaryngoscopy: variability in healthy singers. J Voice 1997; 11:104–107.
24. Casper JK, Brewer DW, Colton RH. Variations in normal human laryngeal anatomy and physiology as viewed fiberscopically. J Voice 1987;1:180–185.
25. Hillel AD. Classifying and diagnosing laryngeal dystonia-are we artists or are we scientists? JAMA Otolaryngol Head Neck Surg 2018;144:666–667.
26. Nawka T, Konerding U. The interrater reliability of stroboscopy evaluations. J Voice 2012;26:812.e1–812.e10.
27. Misono S, Marmor S, Roy N, Mau T, Cohen SM. Multi-institutional study of voice disorders and voice therapy referral. Otolaryngol Head Neck Surg 2016;155:33–41.
28. Francis DO, Daniero JJ, Hovis KL, et al. Voice-related patient-reported outcome measures: a systematic review of instrument development and validation. J Speech Lang Hear Res 2017;60:62–88.
29. Graber ML, Sorensen AV, Biswas J, et al. Developing checklists to prevent diagnostic error in emergency room settings. Diagnosis (Berl) 2014;1: 223–231.
30. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003;78:775–780.
31. Pines JM. Profiles in patient safety : confirmation bias in emergency medicine. Acad Emerg Med 2006;13:90–94.
32. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35:1668–1676.
33. Bonilha HS, Focht KL, Martin-Harris B. Rater methodology for stroboscopy: a systematic review. J Voice 2015;29:101–108.

Appendix A

Brief History A

Patient is a professional singer without history of voice problems or current voice complaints. Patient is concerned about the potential risk of vocal fold injury associated with intubation required for her upcoming elective surgery. Requested referral to obtain baseline videostroboscopy and voice assessment prior to undergoing surgery.

Diagnostic Hypothesis: Normal laryngeal appearance
Unique VLS Expectation: Within normal limits

Brief History B

Sudden onset of mild voice difficulties with intermittent choking with liquids following left hemi-thyroidectomy six weeks ago with gradual improvement in both voice and swallowing symptoms. Patient has minimal occupational voice demands. Medical history is otherwise unremarkable.

Diagnostic Hypothesis: Left vocal fold paresis
Unique VLS Expectation: Vocal fold immobility (left)

Brief History C

Patient with concerns about throat dryness and sensation of lump in the throat as well as mild hoarseness in the morning with normal voice in the afternoon and evening. Patient has minimal voice demands. Admits to eating fast food, spicy foods, and consuming late-night meals frequently. Consumes less than two glasses of water per day and three caffeinated beverages per day.

Diagnostic Hypothesis: Laryngealpharyngeal reflux
Unique VLS Expectation: Abnormal posterior laryngeal appearance

Brief History D

Patient with extensive occupational voice demands who experienced a sudden and severe change in voice quality that improved with voice rest. Patient with continued vocal fatigue and increased effort. Unremarkable medical history.

Diagnostic Hypothesis: Resolving hemorrhage/asymmetric lesion(s)
Unique VLS Expectation: Mucosal wave abnormality (right)

Appendix B

Expert Rating Form

Instructions: Please take as much time as you need to become familiar with the following definitions and the way that we would like you to rate the following laryngoscopic and videostroboscopic parameters prior to viewing the videos. During the rating phase, you can refer to this form and these definitions as frequently as you would like. You can pause and resume playing the videos as often as needed. Each video will include sustained phonation, pitch glides, high pitched phonation, and laryngeal DDKs, which will be repeated in sequence three times. You can play the video in its entirety a total of two times. Although you cannot view the video more than two times, you can take as much time as needed after each video viewing to complete the rating form before beginning the next video rating task.

Gross Laryngeal Structure and Motion

Posterior laryngeal appearance--*findings of erythema, edema, cobblestoning, or other irregularities of the posterior pharyngeal wall, posterior pharyngeal bar, posterior commissure and inter-arytenoid mucosa.*

NONE *MILD* *MODERATE* *SEVERE*

Vocal fold appearance-*gross structural appearance of the tissue of the true vocal folds.*

ERYTHEMA-includes all observations related to the redness of the tissue as it pertains to the coloring of the vocal folds.

RIGHT VOCAL FOLD

NONE MILD MODERATE SEVERE

LEFT VOCAL FOLD

NONE MILD MODERATE SEVERE

LESION- presence of diffuse or discrete lesion involving any aspect of the true vocal fold.

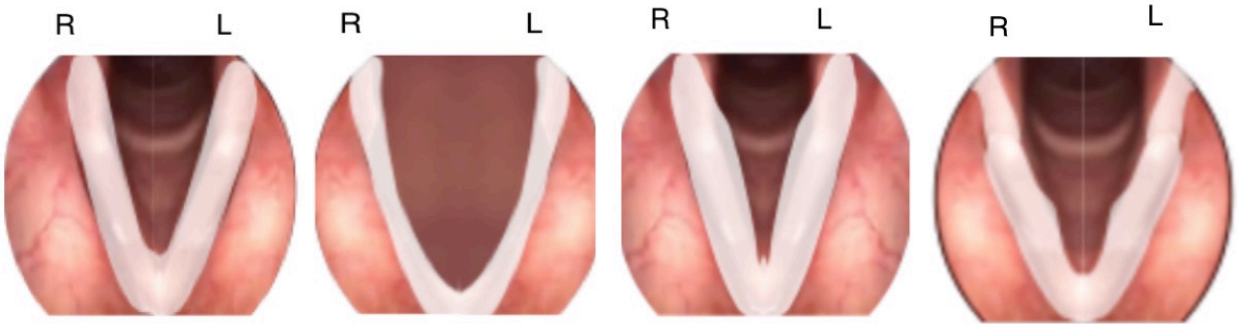
RIGHT VOCAL FOLD

NONE SMALL MODERATE LARGE

LEFT VOCAL FOLD

NONE SMALL MODERATE LARGE

VOCAL FOLD EDGE/CONTOUR--as measured during full abduction of the true vocal folds spanning from the vocal process to anterior commissure. Vocal fold edges are described as ***normal or abnormal.*** *Normal describes a straight vocal fold edge. Abnormalities can include any of the following appearances: concave, convex, or irregular/other.* ***If abnormal, please estimate severity of observed abnormality. Report medial edge appearance independently for the left and right vocal folds.***



Straight Concave Convex Irregular/Other

RIGHT VOCAL FOLD

NORMAL (STRAIGHT)			
ABNORMAL (CONCAVE, CONVEX, IRREGULAR, OTHER)	MILD	MODERATE	SEVERE

LEFT VOCAL FOLD

NORMAL (STRAIGHT)			
ABNORMAL (CONCAVE, CONVEX, IRREGULAR, OTHER)	MILD	MODERATE	SEVERE

VOCAL FOLD MOBILITY: This is rated as the movement of each of the vocal folds toward and away from the midline at the level of the cricoarytenoid joint when producing a laryngeal diadochokinetic task. ***Report vocal fold mobility independently for the left and right vocal folds.***

RIGHT VOCAL FOLD

NORMAL	MILDLY REDUCED	MODERATELY REDUCED	SEVERELY REDUCED/ABSENT

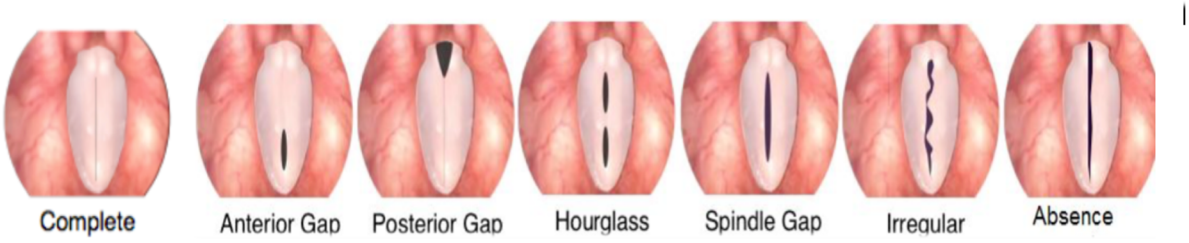
LEFT VOCAL FOLD

NORMAL	MILDLY REDUCED	MODERATELY REDUCED	SEVERELY REDUCED/ABSENT

VIBRATORY CHARACTERISTICS

Evaluation of Vocal Fold Vibratory Characteristics: *Three consecutive videostroboscopic cycles from the middle of a stable phonation should be viewed as a basis for rating the following parameters. Ratings for these parameters should be reported for typical/normal phonation (typical pitch and loudness)*

GLOTTAL CONFIGURATION --The glottal configuration during maximum closure (i.e., closed phase of the glottal cycle) should be rated as:



Complete Closure: Occurs when there is no gap evident on maximal closure.

Anterior Glottal Gap: Occurs when closure is accomplished in the posterior part of the larynx, but a gap remains at some point in the anterior third.

Posterior Gap: Occurs when closure is accomplished along the membranous portion of the glottis, but a gap remains at the posterior glottis. If present, the posterior gap can be of two types: (a) cartilaginous gap only or (2) cartilaginous gap extending into the membranous portion.

Hourglass Gap: Configuration occurs when closure is accomplished somewhere along the membranous portion of the vocal folds, but the gaps are seen both anteriorly and posteriorly to the point of closure.

Spindle-Shaped Gap: Occurs when there is a gap along the membranous portion of the vocal folds with approximation at the vocal processes.
Irregular Glottal Closure: Occurs when the degree of closure varies along the length of the vocal folds. In some places, closure may be complete, whereas in other places, a gap may be observed. The glottal space will not appear as a straight line, but, rather, it will have an irregular contour.

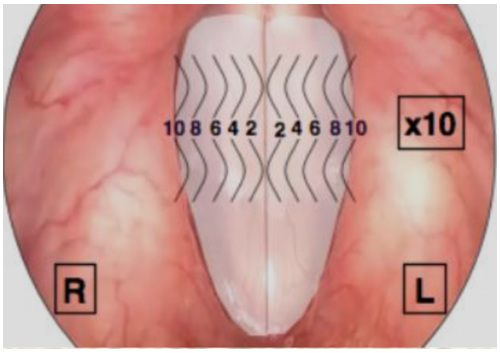
Absence of Closure: A lack of glottal closure exists between the vocal fold along the entire length of the vocal folds, including the cartilaginous portion and the membranous portion during maximal approximation.

Variable Closure: When more than one glottal closure pattern is observed within an examination, the pattern should be rated as variable, and the predominant closure pattern should be identified.

GLOTTAL GAP SIZE - glottal gap size estimated during maximum closure (i.e., closed phase of the glottal cycle) should be rated as ***normal/none*** *or abnormal (glottal gap present).* ***Please estimate size of glottal gap as normal/none, mild, moderate, or severe.***

NORMAL/NONE MILD MODERATE SEVERE

MUCOSAL WAVE: Originates on the medial surface of the vocal folds. It becomes visible during laryngeal videostroboscopy as it travels laterally across the superior surface of the vocal folds from the medial edge. Typically, the wave should travel approximately half (50%) of the visible width of the fold for typical pitch and loudness. ***Determine whether the mucosal wave presence/movement is normal (50%) or abnormal. If it is abnormal, please estimate the degree of abnormality (mild, moderate, severe). Report mucosal wave independently for the left and right vocal folds.***



RIGHT VOCAL FOLD

NORMAL (50%)

MILD MODERATE SEVERE

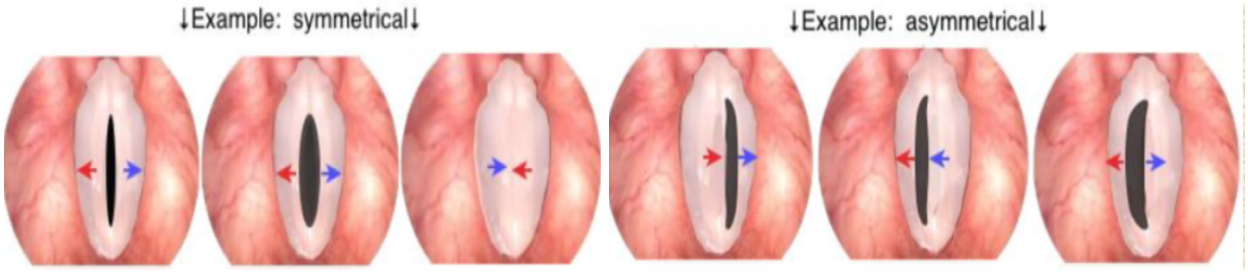
ABNORMAL
(INCREASED OR
DECREASED)

LEFT VOCAL FOLD

NORMAL (50%)

ABNORMAL (INCREASED OR DECREASED) MILD MODERATE SEVERE

PHASE SYMMETRY: Typically, the vocal folds should appear as mirror images of each other in terms of timing of opening, closing, and maximum lateral–medial excursion. ***Rate as normal (consistently symmetric) or abnormal (inconsistently or always asymmetric). If abnormal, report whether phase asymmetry is mild, moderate or severe.***



NORMAL
(ALWAYS
SYMMETRIC)

ABNORMAL MILD MODERATE SEVERE

Please answer the following questions based on the clinical information and video provided.

1. My clinical impression is most consistent with a possible diagnosis of

_____.

Please rate your level of confidence in your clinical impression above.

Really Confident	Confident	Not sure	Not confident	Not really confident

2. The treatment I would most likely recommend is _____.

Please rate your level of confidence in your treatment recommendation above.

Really Confident	Confident	Not sure	Not confident	Not really confident

Chapter 4

The effect of auditory information on visual-perceptual ratings of videolaryngostroboscopy and other clinical judgments.

Introduction

Dysphonia is characterized by altered vocal quality, pitch, loudness, and/or vocal effort and is a common symptom and sign of many different types of voice disorders.[1] Voice disorders are defined by the American Speech-Language-Hearing Association (ASHA) as inappropriate or abnormal voice quality, pitch, or loudness of a person's voice, given the "age, gender, cultural background, or geographic location of the individual".[1] This part of ASHA's definition of a voice disorder particularly emphasizes deviations that might cause attention or distress to a listener.[2(p34)] ASHA further defines a voice disorder as present when the speaker expresses concern about a voice that does not meet daily needs, regardless of the perceptions of others.

Voice disorders have many different etiologies. ASHA broadly characterizes the nature of voice disorders as organic or functional. Organic etiologies are commonly structural (e.g., benign or malignant vocal fold lesions, papilloma, scar), and/or neurogenic (e.g., spasmodic dysphonia, unilateral vocal fold paralysis),[1] whereas functional etiologies are those that arise from improper or inefficient use of the vocal mechanism (e.g., muscle tension dysphonia). Individuals can present with multiple types of voice disorders from within the same category, or from across different categories. For example, a person with a vocal fold paresis (i.e., a neurogenic disorder) may also present with secondary muscle tension dysphonia (i.e., a functional voice disorder) that arises when a person compensates for the paresis. Improved treatment effectiveness and better patient outcomes are associated with efficiently and accurately determining the nature of a voice disorder.[3,4]

Clinicians, such as otolaryngologists (ENT)s and speech-language pathologists (SLP)s perform clinical voice assessments to determine the presence, severity, and nature of dysphonia. Physicians render medical diagnoses related to underlying laryngeal pathology, determine

appropriate surgical and/or medical management strategies (e.g., surgery, medication, referral for voice therapy, etc.), and consider how voice symptoms affect individuals in everyday situations. The purpose of an SLP's evaluation commonly overlaps with those of ENTs. SLPs assess voice production and the psychosocial impact of the voice disorder, determine prognosis for change with behavioral interventions, provide recommendations for treatment, and recommend referrals as needed. Voice assessment is also used to measure treatment outcomes and/or for surveillance. Consequently, voice evaluations require many different clinical assessment tools.

Recently, an ad hoc committee sponsored by the Voice and Voice Disorders Interest Group (ASHA Special Interest Group 3) was charged with developing a minimal set of recommendations for instrumental voice assessment for patients evaluated in the United States. Specifically, the committee recommended a minimal core set of measures to be used in an instrumental voice assessment protocol (IVAP).[5] The IVAP includes videolaryngostroboscopic, acoustic, and aerodynamic measures. Data derived from this protocol are meant to complement other information and measures gathered during a typical evaluation that include a thorough case history, auditory-perceptual voice assessment, and patient-reported outcome measures.[5,6] In a typical voice evaluation, information obtained from these clinical voice assessment tools is combined to determine the presence, severity, and nature of dysphonia, provide patient education, determine prognosis for improvement with various voice treatments, measure treatment outcomes, and understand the impact of voice symptoms on activities of daily living.[6,7] Yet, few studies have examined how clinicians combine information from clinical voice assessment tools, including videolaryngostroboscopy (VLS), to determine the etiology and make other types of clinical decisions.[8–10] One overarching goal of this study was to address this gap, having important implications for clinical practice and future research designs.

Visual-Perceptual Measures of Videolaryngostroboscopy (VLS)

VLS is one of the most common and important clinical voice assessment tools used to evaluate the presence, severity and nature of a voice disorder.[5,6] The IVAP[5] specifies that all individuals with dysphonia be assessed with videoendoscopic measures of laryngeal structure and function, including methods that can be used to evaluate vibratory function, such as VLS. This recommendation is based on evidence from other studies that suggests that VLS is essential for determining the nature of dysphonia, making treatment recommendations, and assessing treatment outcomes.[8,9,11]

While VLS is considered the gold-standard tool for visualizing laryngeal structure and function during voice production, several limitations of the tool may threaten its validity.[7] The validity of a clinical assessment tool means that it measures what it is intended to measure. VLS is used to determine the presence, severity, and nature of a voice disorder. For example, if it is meant to be used to determine the nature of a voice disorder, it should be sensitive (i.e. able to identify abnormalities when they exist), and specific (i.e. able to determine that an abnormality is absent when it does not exist), with acceptable overall diagnostic accuracy (high sensitivity and specificity).[7] Previously, one systematic review by Roy and colleagues[7] identified few high quality studies that could be used to support the validity of different types of voice measures for determining the presence, severity, or nature of a voice disorder. Interestingly, while visual-perceptual judgments of voice using VLS were part of this review, most studies that included a standard reference for comparison used laryngeal imaging to determine the nature of voice disorders. The authors stated that "the consistent use of laryngeal imaging as a standard reference suggests that experts already agree that imaging is critical for determining the presence/absence or nature/etiology of a voice disorder",[7(p219)] although none of the studies included in the

systematic review directly addressed this consideration. Regardless of the circular nature of this issue, Roy and colleagues[7] highlighted the need to develop a broader high quality evidence base to support the validity of voice measures, including VLS.

Sources of Variability in VLS Ratings

There are at least two major challenges to developing a broader high quality evidence base that can be used to support the validity of VLS. One relates to the paucity of gold standards with which to compare the diagnostic accuracy of a tool such as VLS for determining the nature of a voice disorder.[8,12–14] In fact, as described above, most experts would agree that VLS is already considered the gold standard. However, because diagnosis of a condition using VLS is subjective and may vary between clinicians, agreement among experienced clinicians or agreement with one expert laryngologist is often used as a surrogate gold standard for diagnostic accuracy.[9,15,16]

A second difficulty with establishing the validity of VLS is that it is subject to many sources of variability, and these sources are both known and unknown. This problem has been exacerbated by the lack of a theoretical model with which to explain these sources of variability. However, one related conceptual model[17] has been used to explain the many sources of variance associated with mapping an acoustic voice signal onto auditory-perceptual voice quality ratings. This model[17] proposed by Kreiman and colleagues includes listener factors and rating task factors, and the interaction between these factors.

Given the similarities between auditory-perceptual ratings of voice and visual-perceptual ratings of VLS, Kreiman's model may be adapted and extended to provide a systematic framework for explaining and testing factors that affect the validity of VLS (the reader is

referred to figure 3 Chapter 1). First, when clinicians rate a static or moving image of the larynx during phonation with regard to a particular structural or vibratory parameter (e.g., vocal fold edge; mucosal wave), the rating they derive may be affected by *several factors associated with the rater* (i.e., rater factors). Rater factors may include rating experience or professional background, which have been found to affect the reliability of judgments of VLS.[3,9,18]

Second, several *rating task factors* have been pinpointed as sources of variability in VLS interpretation. For example, rating task factors might relate to the stimuli that are included in the rating task, rater training for the task, and the types of rating tools that are used to obtain VLS measures. For example, rating task factors related to the stimuli might include poor representativity of severity or the inherent multi-dimensionality of stimuli that might result in poorer rater reliability.[17,19–21] The amount and type of rater training for the task,[22] and/or the inclusion of anchors [23–25] might also affect rater reliability. Additionally, many different standardized and non-standardized rating tools are used to measure visual-perceptual VLS parameters, both in clinical and research settings.[20,21,26] These rating tools have highly varied response scale characteristics that might also contribute to variability among studies.[26]

A rating tool's response scale characteristics include a scale's visual presentation, "evaluative dimension", polarity, length, metric, and labels.[26,27] For example, it has been hypothesized that by making visual images on the rating tool appear more similar to the VLS recordings being examined, the complexity of the perceptual rating task might be reduced.[28] For example, the Voice-Vibratory Assessment with Laryngeal Imaging tool (VALI)[28] uses this approach: for most VLS parameters, graphics on the rating tool are presented as comparisons from which clinicians make ratings. This approach contrasts with other types of rating scales used for judging auditory- or visual-perceptual stimuli. For example, clinicians might make

ratings using ordinal scales that require severity judgments (normal, mild, moderate, and severe) for the auditory- or visual-parameter under investigation. The type of perceptual parameter (e.g., mucosal wave versus amplitude of vibration) being rated might also affect visual-perceptual VLS ratings.[11,26,28] For example, one systematic review found that structural parameters, such as vocal fold edge, elicited the greatest reliability across rating scales, whereas vibratory parameters, such as amplitude of vibration or mucosal wave, were more difficult to rate.[11]

Many visual-perceptual voice rating tools require clinicians to compare a patient's laryngeal structure and/or function to their own internal standards of what is considered normal or abnormal, which requires relative evaluation.[26] When a standard is needed to provide a relative evaluation, this type of response is categorized as a relative scale measure.[27] For example, visual perceptual voice ratings are performed by comparing a patient's laryngeal structure and/or function to a clinician's internal standard of what is considered (ab)normal based on the perceived age and gender of the speaker. VLS ratings and interpretation are commonly obtained using these relative measures in clinical and research settings.[10,11,29] However, many different metrics and response scale characteristics are used to make visual perceptual ratings of VLS exams.[26]

Some VLS rating tools, such as the VALI,[28] include response scales that are absolute (e.g., amount of excursion), are relative to another standard (e.g., percentage of non-vibrating portion of vocal fold) or relative to time (e.g., percentage of exam time that vocal fold vibration is symmetric).[22,28,30,31,32] These types of ratings do not depend on a clinician's internal standard of what is considered (ab)normal, and might vary less in response to speaker-specific information. It is possible that the relative metric scale used to obtain VLS ratings that depend on a clinician's internal standard for (ab)normal might be more susceptible to bias, compared to

other types of absolute VLS rating scales or metrics, such as those included on the VALI.[28] However, comparisons of different types of response scale characteristics have not yet been examined for VLS ratings. Each of these rating task factors must be considered when developing clinical protocols and designing research studies.

Finally, one rating task factor that has not been systematically investigated relates to information about a speaker that is known prior to and/or during the rating task. While this has received some attention in the area of auditory-perceptual evaluation, very few studies have investigated how speaker-specific information affects ratings and interpretation of VLS.[10,33–35] A summary of these findings is provided next.

The Effect of Speaker-Specific Information on VLS Ratings

One rating task factor that commonly differs in laboratory and clinical studies relates to what clinicians know about a speaker prior to performing auditory and visual-perceptual voice rating tasks. The information that is known about a speaker might provide cues about the presence, severity, or nature of a voice disorder. In a clinical setting, visual-perceptual ratings of VLS are commonly made after case history information is obtained, and clinicians have been exposed to a speaker's voice quality. However, limited or no information about a speaker is known during visual-perceptual rating tasks in a laboratory setting. Few studies have investigated the effect of different types of speaker-specific information (medical laryngeal diagnosis, risk and protective factors) that might also suggest the absence/presence, severity, and/or nature of a voice disorder on perceptual voice measures, including both auditory-perceptual and visual-perceptual VLS measures.[10,33–35]

Many studies have examined the effect of patient information on the detection and/or interpretation of visual-perceptual abnormalities in broader fields of medicine.[36–39] And, a few studies have examined the effect of speaker-specific information on auditory or visual-perceptual voice ratings.[10,33,34] Findings from these studies[10,33,34,36,37,40] commonly attribute the effect of patient information on perceptual ratings to cognitive biases. Cognitive biases are defined as predictable systematic deviations from a normative response which can result in "errors".[41] For example, confirmation bias asserts that examiners might look for, observe, or interpret findings that support a preliminary hypothesis/impression, and ignore findings that refute this hypothesis.[42] Therefore, different types of speaker-specific information might be used to generate hypotheses about the presence, severity, or nature of dysphonia that affect perceptual judgments. To date, the effect of risk and protective factors, referring medical diagnoses, and clinical vignettes intended to suggest a particular voice disorder etiology on perceptual voice tasks have been investigated.[10,33,34,35] Together, findings from these studies provide a basis upon which the effects of other types of speaker-specific information, such as auditory information, might be formed.

Risk/Protective Factors

Risk and protective factors included in a case history might suggest the presence or absence of a voice disorder. For example, Teitler[33] examined differences in perceived VLS severity ratings obtained from 19 ENTs and SLPs when VLS exams were presented without a case history, or after being presented with one of two types of fictional case histories. One fictional case history was associated with risk factors for voice disorders (e.g. smoking, dehydration, voice misuse), whereas the other case history was associated with protective factors

(e.g. hydration, voice conservation). The results showed that in the presence of a case history that suggested a *risk factor* (e.g., this person is a smoker, poor hydration), clinicians rated most VLS exams as more severe than when the same exam was rated with a case history that suggested a *protective factor* (e.g., this person is well hydrated, practices vocal conservation); statistically significant differences were demonstrated in 4 of the 13 VLS parameters. Interestingly, VLS exams rated without case history information were perceived to have *increased* severity of most VLS parameters compared to ratings obtained with case history information suggesting either risk or protective factors. Unfortunately, Teitler[33] did not complete post hoc analysis of the control group that rated the VLS exams without a case history because this group had "greater regularity of rating VLS exams in a clinical setting" compared to other rater groups. Therefore, careful consideration must be given to within- versus between-subjects research study designs to control for these types of clinician factors. Teitler[33] also noted that the effect of fictional histories was greatest for VLS exams that were mild to moderate in severity. It is possible that perceptual judgments might be less stable and more susceptible to the effect of speaker-specific information when making ratings in the mild to moderate range, versus those that are near normal or severe. Therefore, the severity of the stimuli must also be considered in any study that investigates these factors.

Medical Laryngeal Diagnoses

In addition to suggesting the presence of a voice disorder, speaker-specific information might suggest a specific voice disorder etiology. Several studies have found that perceptual voice ratings are influenced by speaker information that suggest the presence and/or nature of a voice disorder, such as a referring medical diagnosis or a case history that is strongly suggestive of a

particular voice disorder etiology.[10,33–35] For example, Sauder and colleagues[10] showed that experienced ENTs and SLPs were more likely to detect and perceive abnormalities as more severe when they were provided case histories that suggested a specific associated voice disorder etiology.

These effects have similarly been shown in auditory-perceptual evaluations. Both Eadie et al.[34] and Sauder and Eadie[35] observed an effect of referring medical laryngeal diagnoses on perceived severity of the auditory-perceptual parameters associated with different medical laryngeal diagnoses (e.g. vocal fold lesions, paralysis/paresis). For example, a priori knowledge of the diagnosis of vocal fold lesions, such as vocal fold nodules, increased the perception of roughness in a speaker's voice compared to judgments of the same speaker without such diagnostic knowledge. Similar to the results found by Teitler[34] for judgments of VLS, the magnitude of the effect of medical laryngeal diagnoses suggesting pathology was also greater when a voice stimulus was mild to moderate, versus normal or severe. These results similarly suggest that severity of the stimulus must be accounted for in examining any speaker-specific effects. Interestingly, Sauder and colleagues[10] found that the effect of the magnitude of differences in perceived roughness and breathiness severity was increased when inaccurate referring medical laryngeal diagnoses versus accurate referring medical laryngeal diagnoses were presented to listeners. Therefore, the accuracy or consistency of the speaker-specific information with the VLS exams must also be considered in any study design used to investigate these factors, as the magnitude of the effect was increased when speaker-specific information was inaccurate.[10]

In addition to *risk/protective factors* or *medical laryngeal diagnoses*, there are other types of speaker-specific information that might also suggest the presence, severity, or nature of a

voice disorder, and could similarly influence VLS ratings and clinical judgments. For example, auditory cues might suggest different levels of dysphonia severity that also indicate the presence or severity of a voice disorder. In addition, specific voice characteristics (e.g. phonatory breaks, rhythmic fluctuations in pitch or loudness) might suggest a particular voice disorder etiology (e.g., neurogenic). Finally, speaker-specific information provided by particular auditory cues might bias VLS ratings. However, the effect of auditory cues has not yet been examined. Thus, the broad goal of this study was to evaluate the effect of the presence of auditory cues suggesting different levels of dysphonia severity on VLS ratings.

Auditory Cues

Clinicians are exposed to auditory cues *prior* to performing VLS ratings in a clinical setting. Auditory cues are also present *during* VLS exams in clinical, and many laboratory settings.[10,22,33,42,43,47] In fact, one limitation of the studies by Teitler[33] and Sauder et al.[10] is that although these studies showed an effect of speaker-specific information on visual-perceptual VLS ratings, clinicians who evaluated the VLS exams did so in the presence of *auditory information*. It is possible that these auditory cues provided clinicians with additional information about a speaker that also suggested the presence, severity, or nature of a voice disorder, which could have affected the results of these studies.

Although auditory information is present when VLS exams are evaluated in clinical settings, auditory information is sometimes omitted in laboratory studies.[5,22,28,29] Auditory information is excluded to provide comparisons with high speed video exams that do not contain auditory information.[5,22] Occasionally, it is excluded with the contention that the presence of this auditory information may also bias visual-perceptual judgments.[28,29] The direction of the effect

of auditory cues on visual-perceptual ratings might be similar to those observed for other types of speaker-specific information. However, its effect has not yet been examined. For example, one hypothesis is that auditory cues that suggest increased dysphonia severity might be associated with *increased* perceived VLS severity[10,34,35] compared to VLS exams rated without auditory cues. The presence of auditory cues that suggest the absence of a voice disorder (i.e., a typical sounding voice) might result in *decreased* VLS severity ratings or *no differences* in VLS severity ratings compared to VLS exams rated without auditory cues. This hypothesis is based on similar studies examining the effects of other types of speaker-specific information.[10,33–35] Hence, the primary aim of this study is to determine if the presence of auditory cues suggesting different levels of dysphonia severity might influence VLS ratings, similar to the effects of other types of speaker-specific information in related studies.

In developing specific hypotheses about the effect of auditory information on VLS ratings in this study, it is important to consider aspects of the research design, such as the selected rating scale and its response characteristics. For example, in previous related studies, Sauder et al.[10] and Teitler[33] used relative scale measures to investigate the effect of speaker-specific information on VLS ratings. Specifically, Teitler[33] used a 6 point ordinal scale and Sauder et al.[10] used a 4 point ordinal rating scale with labels (0 = normal, 1 = mild, 2 = moderate, 3 = severe), which required that clinicians perform these ratings relative to their internal standard for what is considered normal. It is possible that these relative scale measures result in instability in rating, making them more susceptible to bias compared to those that do not require relative evaluation. Yet contrary to this notion, Roy et al.[44] reported strong inter-rater reliability for VLS ratings measured using a 100 mm VAS with endpoints labeled ‘‘normal” and “profoundly abnormal”, also considered a relative scale measure.

As previously mentioned, although many visual-perceptual voice rating tools require clinicians to compare a patient's laryngeal structure and/or function to their own internal standards of what is considered normal, other VLS rating tools (e.g., VALI) include response scales that are absolute (e.g., amount of excursion), are relative to another standard (e.g., percentage of non-vibrating portion of vocal fold), or relative to time (e.g., percentage of exam time vocal fold vibration is symmetric).[22,28,30,31,32] It is possible that the relative metric scales used to obtain VLS ratings in the studies by Teitler[33] and Sauder et al.[10] might have been more susceptible to bias compared to VLS rating scales or metrics used in the VALI.[28] Thus, although the primary aim of this study was to examine the effect of auditory cues that suggest different levels of dysphonia severity on vibratory VLS ratings, we included both VAS rating scales that require relative evaluation, and the VALI[28] that does not include relative scale measures for stroboscopic ratings. Although scale properties were not compared directly in this study, we hypothesized that the magnitude of any effect of auditory information would be increased for ratings made using a VAS versus the VALI because relative evaluation of perceptual parameters is required for the VAS measures. Additionally, this study aimed to examine whether the presence of auditory information might affect interpretation of VLS or other types of clinical decisions about diagnosis or treatment.

Clinical Decisions-Diagnostic Coding and Treatment

Visual-perceptual VLS judgments are considered important for rendering diagnoses and making treatment recommendations during voice assessment. Therefore, it is also essential to determine if any observed effects of rating tasks factors, such as the presence of speaker-specific information, might be clinically meaningful. One method that can be used to evaluate the clinical

meaning from this study might be to examine clinical impressions and treatment recommendations. Similar clinical judgments about diagnosis and/or treatment recommendations are commonly used to evaluate the validity of other clinical assessment tools, including those that also rely on visual-perceptual judgments.[16,45–47]

It is possible that the effects of speaker specific information on clinical diagnosis and treatment might be indirect and/or direct. For example, different types of speaker-specific information might result in variability in the perceived severity of a specific perceptual feature, such as vocal fold motion. This might result in a diagnosis of vocal fold paresis versus paralysis, and increase diagnostic disagreement among clinicians.[48] Sauder et al.[10] found that clinicians rated visual-perceptual abnormalities as more severe when a case history suggested its presence. And, accurate versus inaccurate case histories were associated with improved agreement in clinical impressions among clinicians. Similarly, differences in perceived severity of vocal fold mobility or glottal gap size might also lead to different treatment recommendations, such as a recommendation for behavioral intervention versus medical/surgical intervention.

However, speaker-specific information might also directly and uniquely contribute to decisions about clinical diagnosis and/or treatment for many voice disorders. Further, it is hypothesized that speaker-specific information would be of greatest relative importance for making treatment recommendations, when compared to other types of clinical judgments. For example, Sauder et al.[10] found that when voice symptoms were absent from the case history (control), clinicians were less likely to recommend treatment, irrespective of the VLS exam (i.e. normal versus abnormal laryngeal appearance and/or function) that was presented with the control case history. This finding illustrates that speaker-specific information might not only

influence visual-perceptual judgments, but it might also uniquely contribute to clinical decisions about diagnosis or treatment (see Figure I).

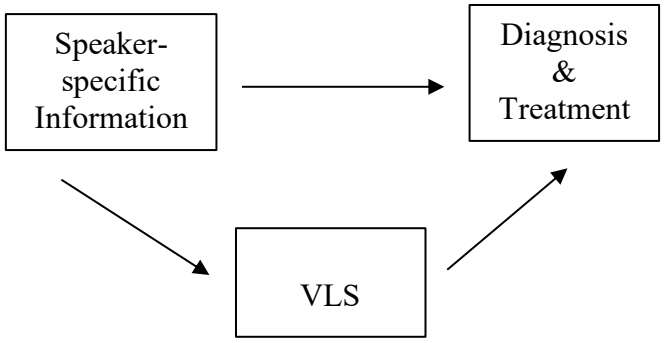


Figure 1 Conceptual model of effects of speaker-specific information on diagnosis and treatment

There are many additional sources of variability in clinical decisions about diagnosis and treatment in a clinical voice setting (e.g., practice patterns, perceived versus actual relative importance of findings from clinical assessment tools, voice disorder diagnosis, etc.).[9] However, obtaining information about clinical diagnosis or treatment is one way to evaluate the overall effects of different types of speaker-specific information on clinical diagnosis and/or treatment, and/or to determine whether any observed differences in perceptual ratings might be clinically meaningful. One caveat is that many decisions about clinical diagnosis and treatment are made "under uncertainty"[5] in laboratory and clinical settings. That is, clinicians analyze information that is often subjective, incomplete, and/or imperfect to make decisions about clinical diagnosis and treatment options.[49] Therefore, measures of confidence are commonly obtained to determine an examiner's beliefs about levels of uncertainty of decisions.[8,16] While several studies have examined the effect of the presence and/or accuracy of speaker-specific information (e.g., medical laryngeal diagnosis, risk and protective factors) on auditory and/or visual-perceptual voice judgments, few have also examined their effect on clinical diagnosis or treatment.[8,10]

Therefore, a secondary aim of this study was to determine the overall effect of this auditory information on diagnostic coding and treatment recommendations.

Methods

Overview

This study used a repeated measures experimental design to address two main objectives. The first aim of this study was to examine the effect of the presence of auditory information (independent variable with two levels: absent or present) on experienced voice clinicians' VLS ratings using a) 100 mm VAS (left endpoint = normal; right endpoint = most extreme example of deviance) for 9 stroboscopic VLS parameters and one measure of overall severity of laryngeal function (dependent variables); and b) the VALI scale for the same 9 stroboscopic VLS parameters (dependent variables) (see Appendix A). The VALI uses a bipolar scale to measure one VLS parameter, phase closure. Therefore, the midpoint of the scale for this VLS parameter measured using the VALI was coded as 0 (nearly equal), positive numbers reflect a predominant closed phase, and negative numbers reflect a predominant open phase. The experimental design also examined the interaction between auditory information and dysphonia severity (normal, mild, moderate-severe) on the dependent variables (stroboscopic parameters derived using a 100 mm VAS and the scale types provided on the VALI and one measure of overall laryngeal function measured using a 100 mm VAS). The second aim of this study was to examine the effect of auditory information on a clinician's exact agreement in diagnostic coding and primary treatment recommendations when VLS exams were viewed with and without auditory information.

In brief, experienced clinicians rated the same VLS exams with and without auditory information that was consistent with the VLS exam presentation (e.g., control VLS exams = no dysphonia, experimental VLS exams = mild to moderately-severe dysphonia), with a 1-2 week washout period between rating sessions. They also provided diagnostic code(s), treatment recommendations, and levels of confidence in these clinical decisions for each VLS exam. Obtaining these data provided some information about how VLS exam findings were interpreted by clinicians. Confidence levels provided one measure of the degree of certainty in recommendations rendered with and without auditory information during VLS interpretation. The repeated measures design was selected to reduce potential confounders associated with rater characteristics. Details related to the clinicians, video stimuli, preparation of the stimuli, and rating procedures are provided below. The recruitment of the clinicians and all procedures used in this study were approved by the Institutional Review Board at the University of Washington.

Clinicians

A power analysis was performed a priori to determine the number of clinicians and videos that would be required to detect an effect of the auditory cues in this study. With anticipated inter-rater reliability coefficients similar to those reported in a previous investigation,[28] a power analysis indicated an 80% chance of detecting a significant difference between conditions (e.g., auditory information present versus absent) at the $p = 0.05$ level of significance with 7 clinicians and 16 VLS exams.

A convenience sample of clinicians was subsequently recruited through a web search and professional contacts. The sample included SLPs who provided voice assessments, regularly interpreted VLS exams, and were employed in the greater Seattle area. While there have not been any investigations examining differences in VLS rating tasks between otolaryngologists

(ENTs) and SLPs, many studies include only SLPs,[22,28,50] including one investigation using the VALI.[28] As a result, to control for possible influence of professional background, only SLPs were included as clinicians in this study.

SLP clinicians were contacted via email, and all individuals: 1) reported greater than 2 years of experience performing voice assessments; 2) reported regular VLS interpretation in their clinical practice; 3) were familiar with the International Classification of Disease 10 (ICD-10) codes; and 4) reported adequate visual acuity to perform voice assessments as part of routine clinical practice, and 5) normal hearing. These inclusion criteria were identified a priori as factors that could affect the reliability of VLS interpretation and/or the validity of performing treatment recommendations. For example, the 2-year experience inclusion criterion was selected to permit comparisons with previous studies and to control for the possible influence of clinical experience on rater-reliability because years of experience and/or frequency of VLS interpretation have shown inconsistent results in previous studies.[29,33] In addition to VLS rating tasks, professional background and/or training has been shown to influence treatment recommendations following voice assessment.[9,26] SLPs do not make medical diagnoses, but they are often required to select diagnostic billing codes to demonstrate the medical necessity of voice assessment and/or treatment for third party payment. Differences associated with training, practice patterns, professional liability, reimbursement, among other factors, might result in dissimilarities in diagnostic coding and treatment recommendations between SLPs and ENTs. Therefore, this study included only experienced SLP voice clinicians who were also familiar with ICD-10 codes used to bill for voice care services in many settings. Finally, all individuals passed hearing screenings at 25 dB SPL for the frequencies of 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz for each ear.

The final sample of clinicians included 8 SLP voice clinicians with 5-20 years of experience (M = 9.6 years; SD = 6.23). Clinicians were all female and between 30 and 59 years of age (M = 41.25 years; SD = 12.44). Most clinicians (87.5%) practiced in outpatient hospital settings, and one clinician practiced in a private practice setting. One clinician used the VALI for rating VLS exams, but others reported no routine clinical use of any specific standardized rating form. Clinicians were paid $15.00 for their participation in this study.

Procedure

VLS Stimuli: Selection and Characteristics

Preliminary control and experimental VLS exams were selected from the University of Washington Speech and Hearing Sciences Database by the first author (CS). All VLS exams in the clinical database were previously recorded using the same instrumentation, a 70 degree KayPENTAX Rigid Laryngoscope, a 9200C processor, 9100B light source and Toshiba 3CCD Camera with mounted microphone, which meet adequate requirements for validly performing VLS judgments of appearance and function, including vibratory function.[5,51] Videostroboscopic recordings were captured at the standard 30 frames per second and were recorded using a standardized protocol. The audio signal was recorded with an electret microphone mounted to the CCD camera at a sampling rate of 22050 Hz and 16 bits per sample.[51] The videos were edited to include one token of sustained vowel production (approximation of /i/) at comfortable pitch and loudness levels repeated three times, and one token of high pitched phonation (approximation of /i/) repeated 3 times, and fully abducted views during inhalation. Each video included, on average, a total of 1 minute of a standardized evaluation, with stroboscopy (1.5 Hz).

To be included in this study, VLS exams needed to meet the following requirements: 1) met adequate video quality based on IVAP[5] recommendations for brightness and resolution; 2) were from adult males and females >18 years of age; 3) included at least three consecutive videostroboscopic glottal cycles during data acquisition; 4) included sustained phonation at a speaker's self-perceived comfortable pitch and loudness and high pitched phonation during 1.5 Hz strobe operating mode; 5) included views of the entire length and width of the vocal folds during phonation at comfortable pitch and loudness levels; and 6) included a fully abducted view to ensure that clinicians could evaluate the VLS parameter, medial edge. Additionally, all VLS exams during self-perceived comfortable pitched productions were within 2 SD of the average fundamental frequency of male/female speakers, as measured from output from the laryngeal microphone, and were within 2 SD of the average intensity across all VLS exams, as determined by the output from the camera mounted microphone during comfortable pitched sustained phonation. The first author selected preliminary control VLS exams that were judged to be normal in structure and function and selected experimental VLS exams that contained a structural abnormality that affected one or more vibratory characteristics.

To be included in the final sample selection, VLS exams needed to meet several additional criteria. Experimental VLS exams: 1) contained structural abnormalities as judged by the same board certified laryngologist blinded to case history information; and 2) had demonstrable dysphonia, as judged by three experienced SLPs. Control VLS exams included in the final sample were judged to be within normal limits for structure and function by a board certified laryngologist (JPG) blinded to case history information; and 2) had no demonstrable dysphonia, as judged by the same three experienced SLPs. Further details about VLS selection criteria as well as audio and video quality are described below.

Selection Criteria Related to Structural Abnormalities

Experimental VLS exams that contained structural abnormalities, as determined by a board certified laryngologist, were selected as experimental VLS exams for several reasons. First, all of the VLS exams in this database were obtained using a rigid endoscope, which is appropriate instrumentation for assessing structural abnormalities. Including only rigid VLS exams helps to control for differences in visual-perceptual judgments related to instrumentation,[52] and controls for the amount and type of auditory information that are present during the VLS exams, as VLS exams obtained using a rigid endoscope contain sustained phonation, but do not contain connected speech samples. Additionally, the accuracy, or consistency of speaker information with the VLS exams is an important consideration. Therefore, we expected a high level of association between the severity of stroboscopic VLS parameters and dysphonia severity in speakers with structural abnormalities, and without significant hyperfunction compared to other voice disorder etiologies.[53,54] Last, there is widespread agreement that visual-perceptual information is important for making clinical decisions about diagnosis and/or treatment, especially in speakers with structural laryngeal abnormalities.[8,9]

Selection criteria related to dysphonia severity

To determine the presence and severity of the dysphonia that examiners would be exposed to during the rating task, voice samples of sustained phonation were extracted from the VLS exams. During VLS data acquisition, the auditory signal was recorded with an electret microphone mounted to the CCD camera. The auditory information was recorded at a sampling rate of 22,050 Hz and 16 bits per second.[51] Recordings obtained using the CCD camera mounted

microphone reduces some variability in mouth to microphone distance. However, the audio input adjustment setting was not recorded for these exams during data acquisition. Because the microphone to equipment distance was also not controlled during data acquisition, noise levels from the equipment was variable between exams.

The audio signal was prepared for a listening task by extracting a 3 second mid-portion of a sustained vowel phonation, produced at a comfortable pitch and loudness level, from the VLS exams. The audio sample was then normalized using Adobe Audition (V11.1.0.184). Voice samples were entered into a custom-made, web-based computer program that generated speaker order, presented rating scales, and recorded responses. Three expert speech-language pathologists with >10 years of experience interpreting VLS exams, and performing auditory-perceptual voice assessment, rated overall voice severity independently. Before performing the rating task, the expert SLPs were provided information about the task and were familiarized with the type of voice stimuli to be rated. Clinicians were instructed that the task included voice samples extracted from the mid portion of sustained vowels during VLS exams. They were also instructed that voice samples were obtained from adult males and females with and without voice disorders.

During the auditory-perceptual rating task, expert SLPs rated overall voice severity using a 100 mm VAS after listening to each voice sample one time via headphones (Sony MDR-7506). These VAS scales were similar to the rating scale used in the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)[55,56] as well as prior studies that included VAS scales to make VLS ratings.[44,57] The left endpoint of this rating scale was marked as ‘‘normal” and the right endpoint was marked as “profoundly abnormal”. Twenty percent of the voice samples were rated twice to assess intra-rater reliability of this measure. Based on ratings from 5 speakers repeated

twice during the rating task, intra-rater reliability was Spearman's rho (ρ) = 0.82. Inter-rater reliability was measured using ICC using a 2 way random effects model, where both clinicians and measures were random effects. For these three experienced clinicians, ICC = 0.87 (95% CI = 0.75-0.93).

Voice samples were then categorized by severity based on ratings obtained during the rating task. Voice samples rated from 0-9 were categorized as normal, 10-29 were categorized as mild, and > 29 were categorized as moderate to severe.[58] Using these cut-off scores for severity, absolute agreement among all 3 clinicians for these categories was achieved for 8 of 28 voice samples. Voice samples that did not achieve consensus using these cut-off scores for dysphonia severity were reviewed in free field and discussed by the consensus panel. After consensus panel discussion, the remaining voice samples were categorized as normal, mild, or moderate-severe. If control and experimental VLS exams were consistent with auditory-perceptual voice ratings (e.g., control VLS exams = normal, experimental VLS exams = mild to severe dysphonia) following auditory-perceptual assessment and consensus panel discussion, they were retained. Because three experimental VLS exams contained auditory information that was perceived as normal, these exams were excluded. Auditory information from the remaining experimental VLS exams were categorized as mildly dysphonic (N=9), or moderate-severely dysphonic (N=5) because few VLS exams that met inclusion criterion (e.g., minimal supraglottic hyperfunction, adequate stroboscopic tracking) were judged to be severely dysphonic. No control exams were excluded based on auditory-perceptual assessment. These experimental and control VLS stimuli were selected to represent a wide range of dysphonia severities to reduce spectrum bias related to the perceptual ratings.

It was expected that the direction and/or magnitude of the effect of speaker specific information on perceptual voice judgements could also be influenced by severity. For example, previous studies[33,34] have shown the strongest effect of speaker-specific information on perceptual judgments for stimuli that were mild (vs. normal or severe), whereas there was no effect of accurate speaker-specific information on control stimuli. Therefore, VLS stimuli that contained auditory cues suggesting mild dysphonia were oversampled to ensure that the design might be sensitive to detecting any such effects.

Final Selection Criteria

Additional characteristics of the 14 experimental VLS stimuli were finally considered. To ensure structural abnormalities were represented on both right and left vocal folds to control for any possible differences in effects due to characteristics of the VLS stimuli, such as sidedness[28,29] (e.g., structural abnormalities occurring with greater frequency on one vocal fold), the medical laryngeal diagnoses and VLS exams were reviewed prior to finalizing the VLS exam selection. To improve the external validity of the study, VLS exams were also reviewed to ensure that both male and female speakers without dysphonia, with mild dysphonia, and moderate-severe dysphonia were included in the final VLS selection.

The final set of VLS exams included in this study were: four control VLS exams (3 females, 1 male) of the highest quality that contained auditory information suggesting normal voice quality; and 12 experimental exams (8 females, 4 males) from speakers with voice complaints and structural abnormalities that affected vibratory characteristics, and who demonstrated a range of dysphonic severities (8 mild; 4 moderate or severe). Details of the selected VLS exams are presented in table 1.

Table I Speaker's sex, medical laryngeal diagnoses, ICD-10 codes, dysphonia severity

Speaker	Sex	Diagnosis	ICD-10 Code	Dysphonia Severity
1	M	acute laryngitis	J04.0	Moderate
2	M	chronic laryngitis	J37.0	Mild
3	M	chronic laryngitis, muscle tension dysphonia	J37.0, R49.0	Mild
4	F	bilateral cyst	J38.3	Moderate
5	F	hemorrhage, cyst	J38.3, J38.7	Mild
6	F	nodules	J38.2	Mild
7	F	papilloma	D14.1	Moderate
8	F	polyp-l, reactive lesion-r	J38.3, J38.1	Moderate
9	F	pseudocyst-r, reactive lesion-l	J38.3, J38.7	Mild
10	F	bilateral pseudocyst, varix-r	J38.3, J38.7	Mild
11	M	pseudocyst-r, vocal fold edema-l	J38.3, J38.4	Mild
12	F	varix-l	J38.7	Mild
13	F	normal	WNL	Normal
14	F	normal	WNL	Normal
15	F	normal	WNL	Normal
16	M	normal	WNL	Normal

M=Male, F=Female, r = right, l = left, WNL=Within Normal Limits

Stimulus Preparation for Rating Procedure

Videos from the final set of VLS exams were edited and saved as AVI files and transferred to a laptop computer (MacBook Pro 13.3-inch (2560 x 1600). They were presented to clinicians using VLC media player (Version 10.5 (928.50) during the rating task (see below)).

Rater Tasks

Questionnaires.

Clinicians completed one short questionnaire to obtain clinician characteristics and to ensure inclusion and exclusion criteria (except hearing screening tests) were met. This questionnaire included information about age, gender, number of years of voice experience,

familiarity interpreting VLS, familiarity and use of ICD codes for billing purpose, practice setting, and information about any standardized rating forms used routinely in clinical practice.

Familiarization task

To ensure adequate rater recruitment and to control attrition due to fatigue, clinicians were not required to complete a lengthy training protocol. Instead, clinicians were first asked to examine the rating form and perform a familiarization task prior to initiating rating at the first rating session. They were also provided an opportunity to repeat the familiarization task prior to the second rating session. During this familiarization task, clinicians watched a video that included the visual images that served as external referents on the VALI.[28] They were also provided operational definitions and rating instructions, in writing, for each VLS parameter. Operational definitions and rating instructions were based on those provided with the VALI,[28] and included in the IVAP.[5] Additionally, clinicians were instructed to make severity ratings for each visual-perceptual parameter using a 100 mm VAS. For example, they selected a percentage of excursion for amplitude of vibration, as instructed using the VALI.[29] Next, they used a 100 mm VAS to rate the degree of deviation from normal of each VLS parameter based on the same operational definition provided by the VALI.[29]

Clinicians were also instructed to rate the overall severity of laryngeal function for each VLS exam. While several methods have been described to determine overall VLS severity,[33,44,53,59] the validity of measures, within and between medical laryngeal diagnoses, is unknown. Therefore, a 100 mm VAS that has been used to assess overall severity of laryngeal structure and function with adequate inter-rater reliability in previous studies[44] was included in this study. Lastly, the familiarization task included an example of VLS exams similar to the

types (normal and structural abnormalities) and ranges of dysphonia severities (normal to moderately-severe) of the VLS exams included in the rating phase. VLS exams included in the familiarization task were not included in the rating phase. This familiarization video task was used instead of providing clinicians with extensive training and visual video anchors for each parameter. Clinicians were able to review the written operational definitions and instructions, and ask questions during the rating task. Prior to beginning the rating phase, clinicians were told that VLS exams included in the rating phase were obtained from speakers with and without voice disorders.

Rating Procedure

Clinicians viewed each VLS exam, made appropriate ratings of VLS parameters, and provided treatment recommendations and medical laryngeal diagnoses using ICD-10 diagnostic billing codes. Clinicians rated all VLS exams either with or without auditory information during the rating procedure. The order that clinicians were provided auditory information with the VLS exams was counterbalanced to reduce order and practice effects. A washout period between VLS ratings of a range of 1-2 weeks was provided to ensure adequate scheduling flexibility, reduce attrition, and rater fatigue.

VLS exam order was randomized for each participant. However, the order that clinicians rated the videos during the first rating session remained the same in the second ratings session. Auditory information during the VLS was presented via headphones (Sony MDR-7506). Headphones were also worn when auditory information was excluded from VLS exams to control for variance associated with ambient noise, or auditory distractions. Clinicians were allowed to view the videos a maximum of three times to control for the number of times viewed.

However, clinicians were able to pause and resume the video, and advance frame by frame, as often as needed to complete the rating forms (see Appendix A). Clinicians received as much time as needed to complete VLS ratings, provide a primary treatment recommendation, and a diagnostic code(s) appropriate for each vocal fold (if different) at the end of each video. Intra-rater reliability was assessed by rating 2 experimental VLS exams (1= mild dysphonia severity, 1 = moderately-severe dysphonia severity) a second time (total videos rated in each session = 12 videos with structural abnormalities + 4 control videos + 2 repeated = 18 videos total). The same two exams were repeated for each rater.

Materials

VLS Rating Parameters and Response Characteristics

VLS ratings of laryngeal appearance, function, and vibratory characteristics were performed using a VLS rating form that includes many of the vibratory and non-vibratory visual-perceptual parameters included on the VALI.[28] The rating form included 9 parameters of vocal function, including non-vibratory and vibratory characteristics: vocal fold free edge contour, supraglottic activity, glottal configuration, mucosal wave, amplitude of vibration, non-vibrating portion, phase closure, phase symmetry, and periodicity (Appendix A). Some parameters were rated for right and left vocal folds, when appropriate. Supraglottic hyperfunction was rated for anterior-posterior and medial-lateral aspects separately. Although all parameters were rated in order to interpret any possible diagnostic codes and treatment related to these VLS features, only vibratory characteristics were included as dependent variables in this study to ensure adequate statistical power (see below).

ICD-10 Codes and Treatment Recommendations

Although SLPs do not make diagnoses, they select diagnostic codes for billing purposes. Clinicians were asked to select the best primary treatment recommendation from a list of the most commonly used codes for voice disorder assessment, including all codes used to describe these VLS exams that were a priori selected by a board certified laryngologist (see Appendix B). Clinicians were instructed that other codes used routinely during clinical practice were also acceptable. Clinicians were also instructed to provide the most specific ICD-10 diagnostic code that was consistent with their clinical impressions for each VLS exam. If clinicians identified different ICD-10 codes for left and right vocal folds, they provided two ICD-10 codes (one for each vocal fold). This level of detail was necessary so that findings, such as the site of lesion (e.g., right versus left vocal fold cyst), could be adequately interpreted. Both treatment recommendations, and the selected diagnostic code(s), provide some information about how clinicians interpreted VLS exams with and without auditory information while controlling for differences in practice patterns between clinicians. Treatment recommendations were rank ordered to reflect the level of care recommended by the clinician (0 = no treatment, 1= surveillance/monitor, 2 = vocal hygiene only, 3 = voice rest/conservation, 4 = behavioral intervention, 5 = medication, 6 = surgical intervention). Clinicians also provided confidence levels associated with these responses (5 = really confident, 4 = confident, 3 = not sure, 2 = not confident, 1 = really not confident; [8] (see Appendix A).

Exit Survey

After completing the second rating session, clinicians completed an exit survey. Questions included: 1) how likely was it that excluding the auditory information from the VLS exam affected VLS ratings?; 2) how likely was it that excluding the auditory information from VLS affected clinical decisions about diagnostic coding?; and 3) how likely was it that excluding the auditory information from the VLS affected treatment recommendations? Responses were obtained using a 100 mm VAS with the left endpoint labeled "not at all likely" and the right endpoint labeled "extremely likely". The responses provided information about how the clinicians perceived the effect of the auditory information on VLS exams without influencing the results of the study.

Statistical Analysis

Effect of Independent Variables on Dependent Variables

Multiple two-tailed paired *t*-tests were used to determine whether there was a main effect of auditory information on VLS ratings, with correction for multiple comparisons to control for any Type I errors. Specifically, post-hoc comparisons were controlled familywise (e.g., for dependent variables on the 100 mm VAS vs dependent variables on the VALI). The interaction between levels of dysphonia severity (none, mild, moderate-severe) and auditory information was also examined using a repeated measures analysis of variance (RMANOVA) to reduce the likelihood of making a Type II error. The primary dependent measures in this study were 9 vibratory VLS parameter ratings using 100mm VAS for each rating, a rating of overall severity of laryngeal function using the 100 mm VAS relative scale, and 9 vibratory VLS parameters using the VALI.

Analysis of Clinical Decisions-Diagnostic Coding and Treatment

Descriptive statistics were used to examine average confidence in decisions about diagnostic coding and/or treatment for all VLS exams. Descriptive statistics were also used to examine clinicians' perceptions about the expected effect of the auditory information on VLS ratings, diagnostic billing codes, and/or treatment recommendations. To determine whether the auditory cues had any effect on diagnostic coding or treatment recommendations, the percent of exact agreement between 1) primary diagnostic billing code for one or both vocal folds and 2) best treatment recommendation when auditory cues were absent versus present were calculated.

Reliability

Intra-rater reliability

Intra-rater reliability was calculated by repeating 2 VLS exams to derive Spearman rho, which does not assume that data are normally distributed. The total percentage exact agreement in a clinician's selection of nominal/categorical data was used to assess intra-rater agreement of nominal/categorical data. Intra-rater reliability for supraglottic hyperfunction was poor when VLS exams were rated with or without auditory information. When auditory cues were absent, intra-rater reliability was highly variable, regardless of whether ratings were obtained using a VAS (Spearman's rho (ρ) = 0. 21-0.92) or the VALI (Spearman's rho (ρ) = 0.08-0.94). When VLS exams with auditory information were provided to clinicians, intra-rater reliability was also variable, regardless of whether ratings were obtained using a VAS (Spearman's rho (ρ) = 0.02-0.83) or the VALI (Spearman's rho (ρ) = 0.04-0.94). Refer to table 2 for intra-rater reliability for VAS and VALI ratings.

Table II Measures of Intra-rater reliability Spearman Rho and % of Exact Agreement for Categorical Variables

Parameter	Auditory Information Absent	Auditory Information Present	Parameter	Auditory Information Absent	Auditory Information Present
VAS-Supraglottic AP	0.44	0.71**	VALI-Supraglottic AP	0.51*	0.72**
VAS-Supraglottic ML	-0.06	0.43	VALI-Supraglottic ML	0.04	0.61*
VAS-Amplitude-R	0.58*	0.52*	VALI-Amplitude-R	0.94**	0.56*
VAS-Amplitude-L	0.70**	0.83**	VALI-Amplitude-L	0.86**	0.57*
VAS-Mucosal Wave-R	0.48	0.46	VALI-Mucosal Wave-R	0.57*	0.43
VAS-Mucosal Wave-L	0.48	0.65**	VALI-Mucosal Wave-L	0.61*	0.45
VAS-Non-vibrating Portion-R	0.53*	0.73**	VALI-Non-vibrating Portion-R	0.27	0.76**
VAS-Non-vibrating Portion-L	0.42	0.79**	VALI-Non-vibrating Portion-L	0.09	0.83**
VAS-Phase Closure	0.66**	0.02	VALI-Phase Closure	0.76**	0.42
VAS-Phase Symmetry	0.88**	0.34	VALI-Phase Symmetry	0.65**	0.31
VAS-Phase Regularity	0.65**	0.38	VALI-Phase Regularity	0.74**	0.34
Overall Severity	0.58*	0.82**		% of Exact Agreement	
VAS Free Edge Contour-Left	0.78**	0.78**	Free Edge Contour-L	81.3	87.5
VAS Edge-Right	0.21	0.73**	Free Edge Contour-R	62.5	87.5
VAS-Glottal Closure	0.92**	0.70**	Glottal Closure	0	12.5

VAS= Visual Analog Scale, VALI=Vibratory Assessment of Laryngeal Imaging, AP=Anterior-Posterior, ML=Medial-Lateral, R=Right, Left

Absolute Agreement in Diagnostic Codes

Absolute agreement in diagnostic coding was calculated by repeating 2 experimental VLS exams when auditory cues were present and when auditory cues were absent. The total percentage exact agreement in a clinician's selection of diagnostic code(s) was used to assess intra-rater agreement in diagnostic coding. Absolute agreement in diagnostic codes when auditory information was absent 10/16 (62.5%) was the same when auditory information was present 10/16 (62.5%).

Absolute Agreement in Treatment Recommendations

Absolute agreement in treatment recommendations was calculated by repeating 2 experimental VLS exams when auditory cues were present and when auditory cues were absent. The total percentage exact agreement in a clinician's selection of treatment recommendations was used to assess intra-rater agreement of treatment recommendations. Absolute agreement was 10/16 (62.5%) for treatment recommendations when auditory information was absent, and was 12/16 (75%) for treatment recommendations when auditory information was present.

Inter-rater reliability

Inter-rater reliability was calculated for all variables measured on ordinal and/or interval levels using intraclass correlation coefficients (ICC) estimates and their 95% confident intervals using SPSS statistical package[60] based on a mean-rating (k = 8), consistency, 2-way random-effects model. This statistic takes into account the effect of rater and video (i.e. two effects) and assumes both are drawn randomly from a larger population.[61,62] ICCs were calculated separately 1) for the VLS with auditory information absent and 2) VLS with auditory information present.

Because it was expected that ICCs might be low for measures of supraglottic hyperfunction and phase regularity based on inclusion and exclusion criteria (e.g., minimal supraglottic hyperfunction, adequate stroboscopic tracking) due to a restricted range of scores,[61] ICCs for these variables were calculated using raw scores and standardized scores for comparison,[62] and were reported only when they differed.

When auditory information was absent, inter-rater reliability for dependent variables measured using a 100 mm VAS scale ranged from ICC = 0.36-0.96. ICCs and their 95% confidence intervals for VLS exams rated without auditory cues using the VAS and the VALI are provided in table 3. Supraglottic hyperfunction-ML and AP demonstrated the lowest levels of inter-rater reliability. Because VLS exams with supraglottic hyperfunction that obscured views of the true vocal fold were excluded, a narrow range of scores for measures of supraglottic hyperfunction were predicted, and ICCs for standardized scores were calculated. However, ICCs and confidence interval width using standardized scores did not vary. Following Portney and Watkin's[63] definitions, the range of ICCs for other VLS parameters measured using a 100 mm VAS scale ranged from 0.65-0.96, indicating moderate to excellent rater reliability, excluding measures of hyperfunction. Similarly, VLS parameters measured using the VALI ranged from ICC = 0.36-0.92 when auditory information was excluded. Inter-rater reliability for measures of supraglottic hyperfunction were also poor when measured using the VALI (Refer to table 2).

When auditory information was present, inter-rater reliability for VLS parameters rated using a 100 mm VAS ranged from ICC = 0.67-0.94, indicating moderate to excellent reliability for all parameters.[63] However, wide confidence intervals were observed for phase symmetry, making interpretation difficult for this VLS parameter. When auditory information was present, inter-rater reliability for VLS exams ranged from ICC = 0.15-0.91 when parameters were

measured using the VALI. Similar to ratings obtained using the 100 mm VAS, wide confidence intervals were observed for phase symmetry. Rater reliability for most vibratory VLS parameters measured using the VALI ranged from moderate to excellent,[63] and were similar to those obtained using the 100 mm VAS. ICCs and their 95% confidence intervals for VLS exams rated with auditory cues using the 100 mm VAS and the VALI are presented in table 3.

Table IIIa Inter-rater reliability for VLS measures obtained for VAS and VALI ratings in the absence of auditory cues

Auditory Cues Absent											
				95% Confidence Interval						95% Confidence Interval	
Parameter-VAS	ICC	df	F Value	Lower Bounds	Upper Bounds	Parameter-VALI	ICC	df	F Value	Lower Bound	Upper Bounds
VAS-Supraglottic AP	0.63*†	15	2.69	0.27	0.85	VALI-Supraglottic AP	0.66*†	15	2.90	0.32	0.86
VAS-Supraglottic ML	0.36†	15	1.55	-0.27	0.74	*VALI-Supraglottic AP Standardized Scores*	0.90*	15	10.33	0.81	0.96
VAS-Amplitude-R	0.86*	15	7.26	0.73	0.94	VALI-Supraglottic ML	0.36†	15	1.56	-0.26	0.74
VAS-Amplitude-L	0.90*	15	10.05	0.81	0.96	*VALI-Supraglottic ML Standardized Scores*	0.41†	15	1.68	-0.17	0.76
VAS-Mucosal Wave-R	0.92*	15	12.21	0.84	0.97	VALI-Amplitude-R	0.86*	15	7.26	0.73	0.94
VAS-Mucosal Wave-L	0.94*	15	15.81	0.88	0.97	VALI-Amplitude-L	0.75*	15	4.05	0.52	0.90
VAS-Non-vibrating portion-R	0.87*	15	7.47	0.74	0.95	VALI-Mucosal Wave-R	0.83*	15	5.71	0.66	0.93
VAS-Non-vibrating portion-L	0.92*	15	12.99	0.85	0.97	VALI-Mucosal Wave-L	0.71*	15	3.44	0.43	0.88
VAS-Phase Closure	0.87*	15	7.61	0.74	0.95	VALI-Non-vibrating portion-R	0.45*†	15	7.63	0.27	0.69
VAS-Phase Symmetry	0.82*	15	5.62	0.65	0.93	VALI-Non-vibrating portion-L	0.92*	15	12.99	0.85	0.97
VAS-Phase Regularity	0.90*	15	10.33	0.81	0.96	VALI-Phase Closure	0.74*	15	3.91	0.50	0.90
VAS Edge-R	0.95*	15	21.86	0.91	0.98	VALI-Phase Symmetry	0.82*	15	5.52	0.65	0.93
VAS Edge-L	0.96*	15	24.11	0.92	0.98	VALI-Phase Regularity	0.88*	15	8.56	0.77	0.95
VAS-Glottic Closure	0.94*	15	17.74	0.89	0.98						
Overall Severity	0.95*	15	21.77	0.91	0.98						

*significant correlations $p < 0.05$

† wide confidence intervals not interpretable

VAS = Visual Analog Scale, AP = Anterior-Posterior, ML = Medial-Lateral, R = Right, L = Left

Table IIIb Inter-rater reliability for VLS measures obtained for VAS and VALI ratings in the presence of auditory cues

Auditory Cues Present											
				95% Confidence Interval						95% Confidence Interval	
Parameter-VAS	ICC	df	F Value	Lower Bound	Upper Bound	Parameter-VALI	ICC	df	F Value	Lower Bound	Upper Bound
VAS-Supraglottic AP	0.84*	15	6.37	0.69	0.94	VALI-Supraglottic AP	0.67*†	15	2.99	0.34	0.87
VAS-Supraglottic ML	0.79*	15	4.64	0.58	0.91	*VALI-Supraglottic AP Standardized Scores*	0.85*	15	6.68	0.71	0.94
VAS-Amplitude-R	0.82*	15	5.61	0.65	0.93	VALI-Supraglottic ML	0.27†	15	1.37	-0.43	0.71
VAS-Amplitude-L	0.79*	15	4.80	0.59	0.92	*VALI-Supraglottic ML Standardized Scores*	0.50*†	15	1.96	0.01	0.80
VAS-Mucosal Wave-R	0.90*	15	9.58	0.80	0.96	VALI-Amplitude-R	0.82*	15	5.61	0.65	0.93
VAS-Mucosal Wave-L	0.93*	15	13.86	0.86	0.97	VALI-Amplitude-L	0.76*	15	4.19	0.53	0.90
VAS-Non-vibrating portion-R	0.89*	15	9.41	0.79	0.96	VALI-Mucosal Wave-R	0.91*	15	10.47	0.81	0.96
VAS-Non-vibrating portion-L	0.91*	15	11.17	0.82	0.96	VALI-Mucosal Wave-L	0.9*	15	9.98	0.80	0.96
VAS-Phase Closure	0.79*	14	4.86	0.59	0.92	VALI-Non-vibrating portion-R	0.82*	15	5.51	0.64	0.93
VAS-Phase Symmetry	0.67*†	15	3.01	0.35	0.87	VALI-Non-vibrating portion-L	0.91*	15	11.04	0.82	0.96
VAS-Phase Regularity	0.85*	15	6.86	0.71	0.94	VALI-Phase Closure	0.79*	15	4.76	0.59	0.92
VAS Free Edge Contour-L	0.94*	15	17.53	0.89	0.98	VALI-Phase Symmetry	0.63*†	15	2.70	0.27	0.85
VAS Free Edge Contour-R	0.93*	15	15.09	0.87	0.97	VALI-Phase Regularity	0.15†	15	1.17	-0.68	0.65
VAS-Glottic Closure	0.92*	15	12.68	0.85	0.97	*VALI-Phase Regularity Standardized Scores*	0.77	15	4.43	0.56	0.91
Overall Severity	0.94*	15	16.63	0.88	0.98						

*significant correlations $p < 0.05$

† wide confidence intervals not interpretable

VAS = Visual Analog Scale, AP = Anterior-posterior, ML = Medial Lateral, R = Right, L = Left

To assess inter-rater reliability for nominal/categorical data (e.g., VALI ratings for glottal closure, free edge contour-right, free edge contour-left), Fleiss Kappa, an extension of Cohen's Kappa was calculated.[64] The Fleiss' kappa statistic measures the reliability of ratings for a fixed number of clinicians who are assigning their observations to categories. Fleiss Kappa was used to assess the agreement in nominal/categorical data. Inter-rater reliability of these data (free edge left and right; glottal closure) measured using the VALI without, and with auditory cues during the VLS rating task ranged from kappa = 0.37-0.59 and kappa = 0.44-0.59 respectively, indicating fair to moderate inter-rater reliability for these VLS parameters following Landis and Koch's guidelines.[64] The kappa values for these VLS parameters are presented in table 4.

Table IV Inter-rater reliability for VALI VLS parameters with nominal data using Fleiss' Kappa (κ) when auditory cues were absent and present

Parameter-VALI	Auditory Cues Absent		Auditory Cues Present	
	Kappa	Strength	Kappa	Strength
Medial Edge Contour-L	0.37	Fair	0.44	Moderate
Medial Edge Contour-R	0.42	Moderate	0.51	Moderate
Glottal Closure	0.59	Moderate	0.59	Moderate

R = Right, L = Left

Results

Effect of Auditory Cues on VAS ratings

A 2 (auditory information: absent or present; fixed, within-subjects) by 3 (dysphonia severity: no dysphonia, mild dysphonia, moderate-severe dysphonia; fixed, between-subjects)

repeated measures analysis of variance (RMANOVA) was used to assess the main effect of auditory cues and its interaction with dysphonia severity on 9 VLS severity ratings using 100 mm VAS, and one rating of overall severity of laryngeal function using a 100 mm VAS. A Dunn Sidak correction was used to control for multiple comparisons. There was no main effect of auditory information on any VLS parameter, F tests, ps > 0.05.

There were no significant interactions between auditory cues (absent or present) and dysphonia severity (none, mild, moderate-severe) observed in this study. Because the effect of auditory information suggesting no laryngeal pathology versus laryngeal pathology was expected to differ, marginal means for control and experimental VLS exams were examined separately. Examination of the marginal means showed that all dependent variables for experimental VLS exams (suggesting mild or moderate-severe dysphonia) were rated as less severe (0.17–2.42 mm) when auditory cues were present compared to absent, but these differences in severity were not statistically significant (see Figure 2a).

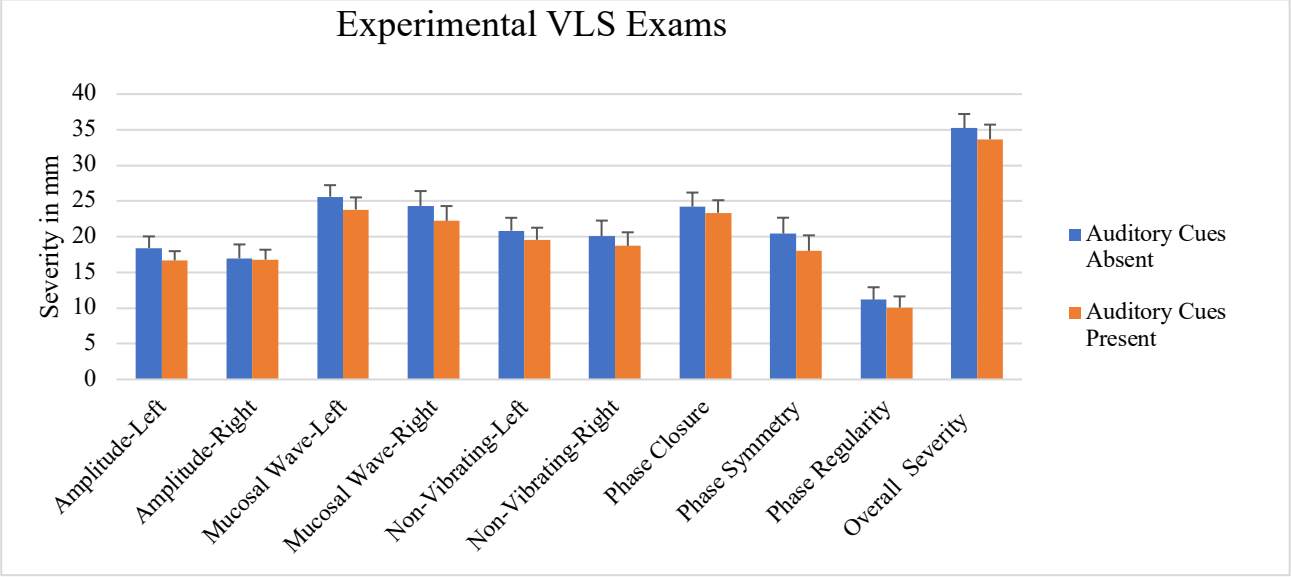


Figure IIa Effect of auditory cues on VLS ratings obtained using the VAS for experimental VLS exams

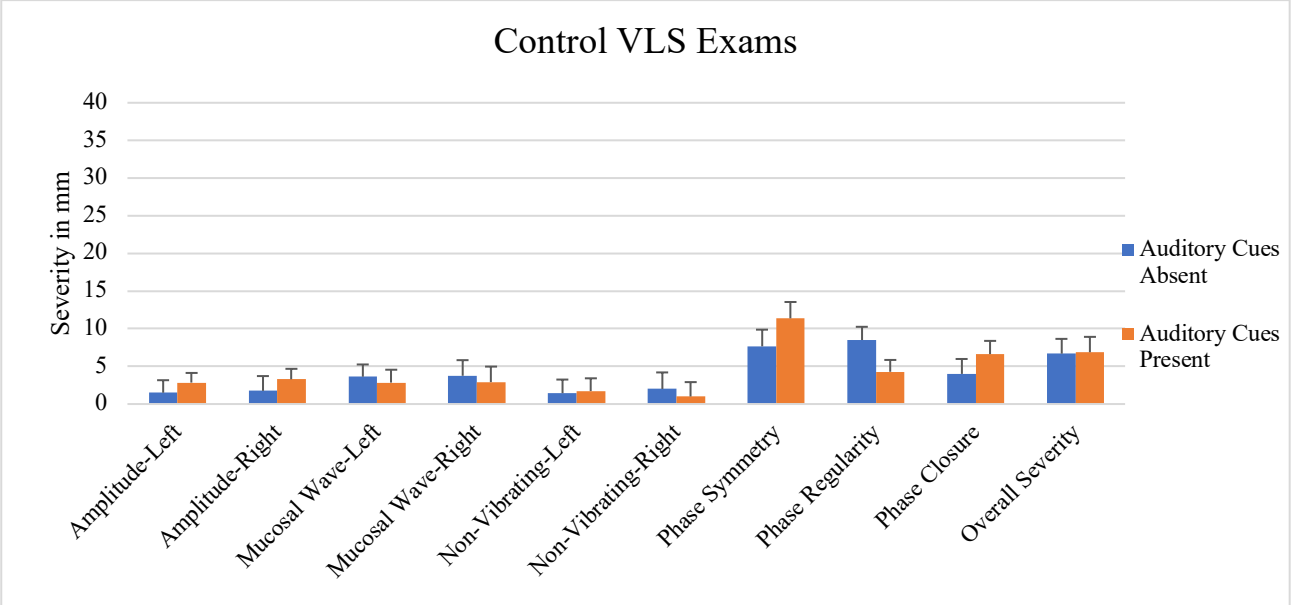


Figure IIb Effect of auditory cues on VLS ratings obtained using the VAS for control VLS exams

Ratings obtained from individual participants for experimental exams were also examined. Although mean group differences suggested decreases in overall average VLS severity when auditory cues were provided, individual data showed that individual clinicians demonstrated both increases and decreases in ratings of experimental VLS exams when auditory cues were present versus absent. The reader is referred to figure 2c for examples of individual clinician ratings for overall severity and a select VLS parameter.

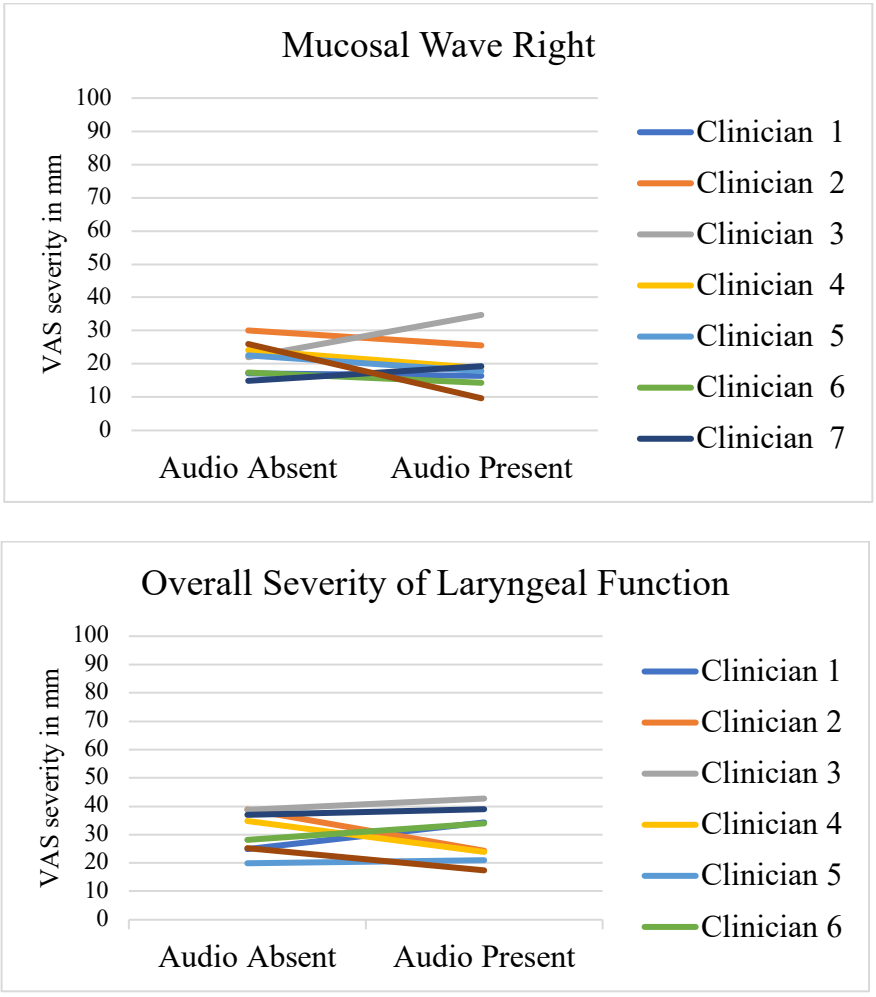


Figure 2c Individual data for select clinician VLS ratings of experimental VLS exams in the absence and presence of auditory cues

Finally, although there were no significant interactions between auditory information and dysphonia severity measured using the VAS, there was a main effect of dysphonia severity on all VLS parameters and overall impressions of laryngeal function, *F* tests, *p*s < 0.05. Post-hoc analysis of dysphonia revealed that differences in marginal means between all levels of dysphonia severity (no dysphonia, mild dysphonia, moderate-severe) were statistically significant (*t*-test adjusted *p*s < 0.05) for most VLS parameters, except phase symmetry and non-vibrating portion-l. For phase symmetry and non-vibrating portion-l, there were only statistically significant differences in marginal means between VLS exams categorized as no dysphonia/control and moderate-severe dysphonia. Differences between all levels of dysphonia severity for overall impressions of laryngeal function also were significant (*t*-test adjusted *p*s < 0.05). In other words, as the speakers' dysphonia increased, there was a similar increase in the average severity of VLS ratings for most VLS parameters as well as overall severity of laryngeal function. These findings corroborated the initial selection of the stimuli.

Effect of Auditory Cues on VALI ratings

A 2 (auditory information: present or absent; fixed, within-subjects) by 3 (dysphonia severity: no dysphonia, mild dysphonia, moderate-severe dysphonia; fixed: between-subjects) repeated measures analysis of variance (RMANOVA) was used to assess the main effect of auditory cues and its interaction with dysphonia severity on 9 VLS ratings obtained using the VALI. A Dunn Sidak correction was used to control for multiple comparisons. For VLS ratings obtained using the VALI, there was a main effect of auditory information on only one VLS parameter, non-vibrating-left(l) vocal fold, $F(9, 128) = 117.18$, $p = 0.05$ (see figure 3a). Examination of the marginal means for auditory information collapsed across all levels of

dysphonia showed that when auditory information was present, on average, VLS ratings for non-vibrating portions-l were increased (M = 1.7%; SD = 0.87%) compared to when auditory information was absent (t test adjusted $p < 0.05$). Because the main effect of auditory information suggesting no laryngeal pathology versus laryngeal pathology was expected to differ, marginal means for control and experimental VLS exams were also examined separately for all VLS parameters. Examination of the marginal means for experimental VLS exams (suggesting mild or moderate-severe dysphonia) and control exams (suggesting no dysphonia) showed no trends across all VLS parameters when auditory cues were absent versus present (see Figure 3a and 3b).

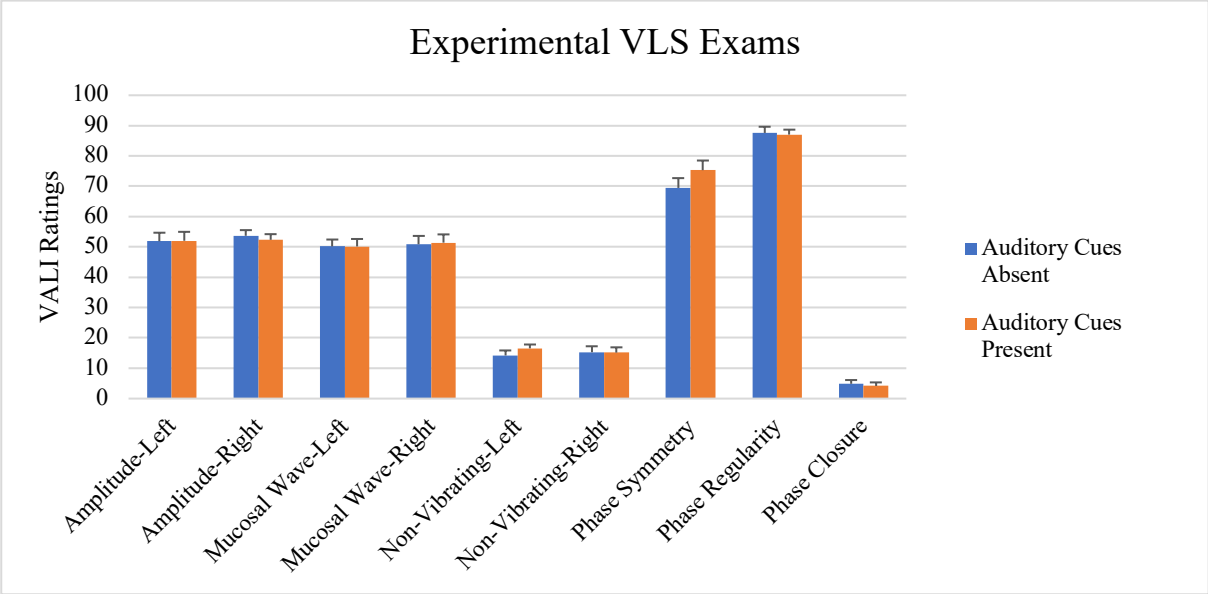


Figure IIIa Effect of auditory cues on VLS ratings obtained using the VALI for experimental VLS exams

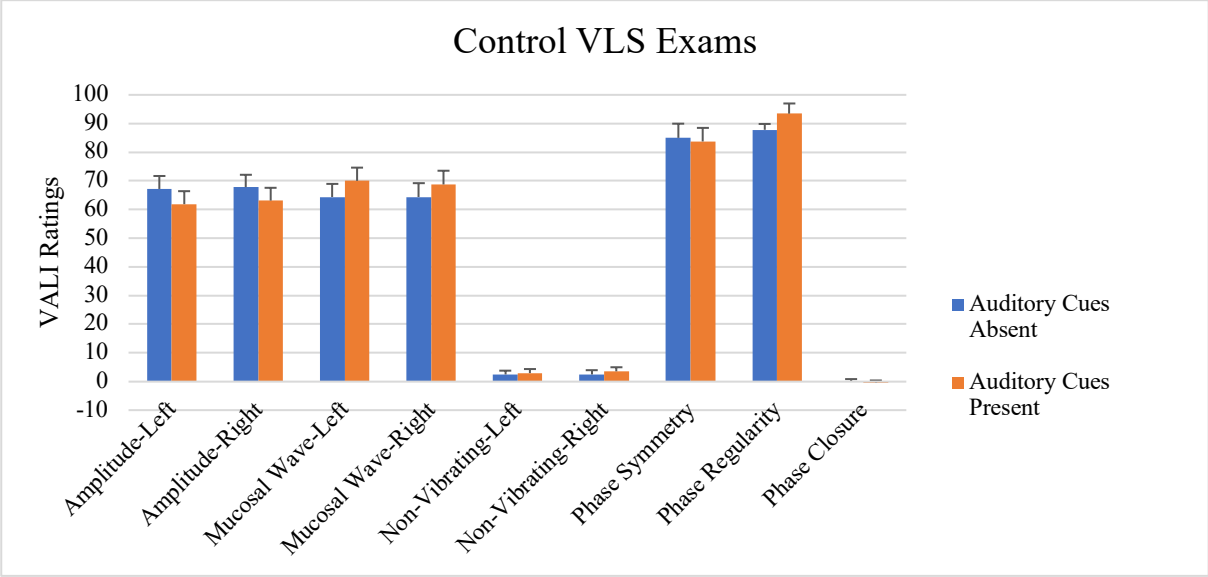


Figure IIIb Effect of auditory information on VLS ratings obtained using the VALI for control VLS exams

In addition to the main effect of auditory information, there was also a significant interaction between auditory cues and dysphonia severity for non-vibrating portion-l (*t*-test adjusted $p < 0.05$) indicating that differences in average VLS ratings for non-vibrating portion-l were also dependent on dysphonia severity. To better understand the nature of the interaction, the group mean for each combination of the two factors was plotted (see Figure 4). As shown in the illustration, the interaction is ordinal, indicating that the differences in VLS ratings for this parameter when auditory information was absent versus present, became larger as auditory information suggested increased dysphonia severity. Simple effects pairwise *t*-tests between the auditory cues for each dysphonia level separately (*p* values adjusted using Dunn-Sidak) showed that the effect of auditory cues on VLS non-vibrating-l was statistically significant when dysphonia severity was moderate-to severe (*t*-test adjusted $p < 0.05$), but not when auditory cues suggested mild or no dysphonia. Differences for non-vibrating-l for speakers with moderate-severe dysphonia when auditory cues were absent versus present were (M = -4.8%; SD = 1.65%).

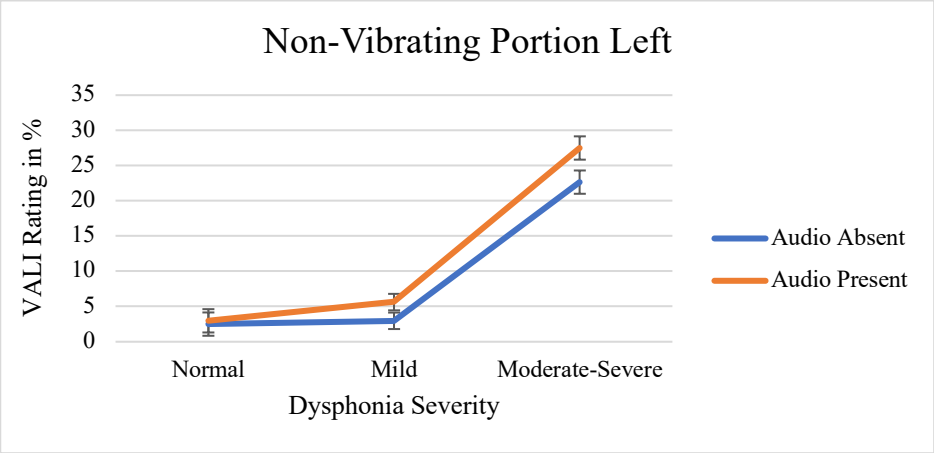


Figure IV Interaction between auditory cues and dysphonia severity for non-vibrating portion left measured using the VALI

Although there was only one significant interaction between auditory information and dysphonia severity using the VALI, there was a main effect of average dysphonia severity for all VLS parameters except phase symmetry, F tests, ps < 0.05. Statistically significant differences between control VLS exams and exams judged to have auditory cues that suggested severe dysphonia were observed for all VLS parameters (t-test adjusted $p < 0.05$), except amplitude-r and phase symmetry. There were no statistically significant differences in mean VLS ratings for control VLS exams versus VLS exams that contained auditory information suggesting mild dysphonia for any VLS parameter (t-test adjusted $p > 0.05$).

Agreement in Clinical Decisions

Diagnostic Codes

Each clinician generated diagnostic billing codes based solely on the information that was provided during the rating task. In order to interpret the findings from VLS ratings, each clinician was able to provide separate diagnostic billing codes for right and left vocal folds if they differed, and/or determine if vocal fold structure and function (right/left) were within normal limits (WNL). Among clinicians, agreement in diagnostic coding mode for individual VLS exams for each vocal fold ranged from 25% to 100%. Because differences in practice patterns can result in variability in diagnostic codes among clinicians, we examined each clinician's level of absolute agreement in diagnostic coding when auditory cues were present versus absent.

Overall agreement in diagnostic billing codes for the same 16 VLS exams rated by each clinician across the two auditory conditions was 93/128 (73%) of VLS exams for at least one vocal fold. When there was disagreement in diagnostic coding across the two auditory conditions

for individual clinicians (representing 35/128 or 27% of VLS exams), the source of these disagreements was commonly 31/35 (89%) related to the presence (e.g., nodules versus WNL) or nature (e.g., sulcus vocalis versus acute laryngitis) of abnormality. Disagreements for a few of these VLS exams 4/35 (11%) did not pertain to differences in the perceived presence/absence of structural abnormalities (e.g., WNL versus dysphonia, WNL vs muscle tension dysphonia).

The overall percentages of observed disagreements in diagnostic coding for one or more vocal folds when auditory cues were absent versus present was similar for experimental VLS exams (27/96, or 28%) and control VLS exams (8/32, or 25%). Because the effect of auditory information might differ for control VLS exams and experimental VLS exams, the types of diagnostic coding disagreements were examined separately. The source of diagnostic disagreement for experimental VLS exams was always related to the presence or nature of structural abnormalities (e.g., WNL versus edema; scar versus benign neoplasm), but there were no clear patterns of change in diagnostic codes associated with the presence or absence of auditory cues. However, results showed that control VLS exams presented without auditory cues were commonly coded as WNL (75%) but were changed to diagnoses such as vocal fold paresis, dysphonia, or edema when auditory cues were present.

Treatment Recommendations

Overall exact agreement about the mode of level of care recommended among all clinicians for each VLS exam ranged from 25% to 100%. To control for difference in practice patterns among clinicians that might contribute to variability in treatment recommendations, we examined agreement in treatment recommendations when VLS exams were evaluated with auditory cues present versus absent for individual clinicians. Overall agreement in treatment

recommendations for the same 16 VLS exams presented with and without auditory information for each clinician was 84/128 (65.6%). The source of these instances of the disagreements in treatment recommendations 44/128 (34.4%) was evaluated to determine whether they might suggest a different level of care. We also examined whether or not any changes in the level of care might be more clinically meaningful (e.g., no treatment versus behavioral intervention) or less clinically meaningful (e.g., voice conservation versus surveillance) because they would potentially affect treatment status. Recommendations for higher levels of care were observed 24/44 (54.5%) and lower levels of care were observed 20/44 (45.5%) when auditory cues were absent versus present. When disagreements were observed, 38/44 (86.4%) of these differences in the level of care that was recommended would have influenced recommendations for treatment versus no treatment (e.g., surveillance versus behavioral intervention) and 6/44 (13.6%) would not have influenced recommendations for treatment versus no treatment (e.g., surveillance versus no treatment).

The overall percentages of observed disagreements in treatment recommendations when auditory cues were absent versus present was slightly higher for experimental VLS exams 44/96 (45.8%) versus control VLS exams 12/32 (37.5%). For experimental VLS exams, the majority of disagreements favored recommendations for higher levels of care 20/32 (62.5%) when auditory cues were absent versus lower levels of care 8/32 (25%) when auditory cues were present. When there were differences in the level of care being recommended (higher or lower) when auditory cues were absent versus present, these differences commonly affected treatment status 19/32 (59.4%), and would not have affected treatment status for 13/32 (40.6%). For control VLS exams, differences favored recommendations for lower levels of care (6/12) 50% when auditory cues were absent versus present. When disagreements about the level of care recommended

(higher or lower levels) were observed, 9/12 (75%) of differences in treatment recommendations would potentially have affected treatment status (e.g., treatment versus no treatment). In other words, when disagreements about treatment recommendations were observed, clinicians were more likely to recommend a higher level of care for experimental VLS exams and a lower level of care for control VLS exams when the auditory information was absent versus present. For both experimental and control VLS exams, the observed differences in the level of care recommended would have commonly impacted recommendations for treatment versus no treatment.

Confidence Ratings:

Average confidence ratings in both diagnostic coding and treatment recommendations for all VLS exam were the same when auditory cues were absent versus present (*M*s = 3.8; *SD*s = 0.58-0.79). In case the results differed for experimental and control VLS exams, average confidence in diagnostic coding and treatment recommendations were examined separately. For experimental VLS exams, average confidence in diagnostic coding when auditory cues were absent and present were the same (*M*s= 3.7; *SD*s = 0.64-0.70). For experimental VLS exams, average confidence in treatment recommendations when auditory cues were absent was (*M* = 3.7; *SD* = 0.70) and when they were present was (*M* = 3.8; *SD* = 0.56). For control VLS exams, average confidence in diagnostic coding when auditory cues were absent or present were the same (*M*s = 3.9; *SDs* = 0.54-0.96). For control VLS exams, average confidence in treatment recommendations when auditory cues were absent (*M* = 3.9; *SD* = 0.76) and when they were present (*M* = 3.8; *SD* = 0.63) were also similar. This finding suggested that average confidence

in diagnostic coding and treatment recommendations did not appear to vary with the presence of auditory information for experimental or control VLS exams.

Exit Survey

Clinicians were also surveyed about their perceptions of the effect of auditory information on visual-perceptual ratings, diagnostic codes, and treatment recommendations using a 100 mm VAS (0 = "not at all likely" and 100 = "extremely likely") upon completing the study. On average, the likelihood that clinicians thought that exclusion of auditory information affected VLS ratings was (M = 53.00; SD = 35.86). Similarly, the likelihood that clinicians thought exclusion of auditory information affected clinical decisions about diagnostic coding was (M = 52.00; SD = 25.38). However, the likelihood that clinicians thought excluding auditory information affected clinical decisions about treatment recommendations was higher, on average, (M = 72.13; SD = 29.11). In other words, on average, clinicians thought that lack of auditory information would be neutral in its effect on VLS ratings and diagnostic coding, but that it was more likely to have affected treatment recommendations. However, there was a wide range of scores observed for these questions during the exit survey.

Discussion

VLS is considered one of the most important clinical voice assessment tools.[5] It is used to determine the presence, severity and/or nature of the voice disorder for speakers with dysphonia.[15] Although VLS ratings are commonly made in the presence of auditory information in clinical settings, auditory information is often extracted in laboratory settings, sometimes with the contention that auditory cues might bias VLS ratings.[28] It was hypothesized that the presence

of auditory information suggesting the absence/presence or severity of a voice disorder might have similar effects to previous investigations examining other types of speaker-specific information. Specifically, it was hypothesized that auditory information suggesting increased dysphonia severity would be associated with systematic increases in vibratory VLS severity ratings for patients with structural abnormalities, similar to the effect of other types of speaker-specific information on auditory-perceptual voice ratings[34,35] and visual-perceptual VLS rating tasks.[10] It was also hypothesized that auditory cues suggesting no dysphonia might be associated with improved VLS ratings or have no effect on VLS ratings, also similar to findings from these studies.[34,35] The magnitude of any effects of auditory information was also hypothesized to be increased for ratings made using a VAS versus the VALI because relative evaluation of perceptual parameters was required for the VAS measures included in this study. The magnitude of any effects of auditory information on VLS ratings was also expected to be greatest for ratings in the mild to moderate severity range because this range demonstrated more susceptibility to bias in two studies.[33,34] Lastly, the effect of auditory cues was expected to be greater for making clinical judgments about treatment recommendations compared to VLS ratings or diagnostic codes due to the possible increased relative importance of auditory information in making these types of clinical judgments.

Effect of Auditory Information on VLS Ratings using a VAS

In this study, there was no significant effect of auditory information on VLS parameters rated using the VAS. However, there were non-significant systematic *decreases* in average perceived VLS severity for all vibratory parameters and measures of overall severity of laryngeal function for experimental VLS exams rated when auditory cues were absent versus present (see

Figure 2a). This finding was similar to Teitler[33] who also found decreased perceived VLS severity for many VLS parameters when case histories suggested risk factors in comparison to no case history, but attributed this finding to differences clinical experience between clinician rater groups. The direction of change observed in this study was contrary to our preliminary hypothesis, and differed from findings from two previous studies examining the effect of medical laryngeal diagnoses suggesting pathology versus no diagnostic information on auditory-perceptual voice ratings.[34,35] Findings of systematic decreases in average perceived severity of all VLS parameters and overall severity of laryngeal function for experimental VLS when measured using VAS might be attributable to examiner bias. Additionally, it was anticipated that vibratory VLS parameters commonly reported to have poor inter-rater reliability and/or wide confidence intervals[11,29,30] might be more susceptible to biases. However, our data showed that inter-rater reliability for nearly all VLS parameters was improved compared to prior studies[29,30] that included a variety of rating scales and tools (see table 3a and 3b). VLS parameters with lower levels of inter-rater reliability were primarily attributed to a reduced range of severity due to VLS selection criteria.

The magnitude of differences in perceived severity for experimental and control VLS exams rated with auditory information absent versus present was also examined. Due to differences in the scales and metrics used to rate different VLS parameters,[10,33] it was not possible to directly compare the small observed differences in VLS severity in this study with previous studies. However, a 100 mm VAS scale similar to the one used to measure VLS severity in this study was used to measure the effect of accurate referring medical laryngeal diagnoses of benign vocal fold lesions or vocal fold paralysis/paresis on ratings of roughness and breathiness in one prior study.[34] Compared to the average differences in *auditory-perceptual*

severity ratings (5-9 mm) when medical laryngeal diagnoses were absent versus present, the magnitude of the observed differences in VLS ratings associated with the presence of auditory information in this study (< 2.5 mm) were reduced compared to this prior investigation.[34,35] While auditory-perceptual ratings and VLS ratings may both be indicators of voice severity, it is still unknown how or whether these types of judgments may be directly compared, despite the use of similar types of scales and units.

This study found no effect of auditory cues on control VLS exams, similar to results from prior studies in which there was no effect of information indicating "normal larynx" on auditory-perceptual voice ratings.[34,35] Although several studies have reported decreases in auditory-perceptual and visual-perceptual VLS severity ratings when speaker-specific information suggested the absence of a voice disorder,[10,33,35] these studies included fictional speaker-specific information and/or inaccurate speaker-specific information that was inconsistent with the stimulus. The potential "accuracy" or consistency of the speaker-specific information has been shown to influence the magnitude and direction of the effect of speaker-specific information on perceptual judgments in voice assessment and in broader fields of medicine.[36,38] Although the presence of inaccurate speaker-specific information is also common scenario in clinical practice,[9] this study attempted to control for the "accuracy" of the auditory information. The accuracy of the auditory information will be discussed further below.

One additional preliminary hypothesis in this study relates to a response scale characteristic, relative versus absolute measures, that differs for vibratory VLS parameters rated using the VAS versus the VALI. Although scale type was not included as an independent variable in this study, we expected to observe increased magnitude of any effects of auditory information for VLS ratings using a VAS versus the VALI. This hypothesis was based on the

fact that relative evaluation of perceptual parameters was required for the VAS measures included in this study.

Effect of Auditory Information on VLS Ratings using the VALI

It was hypothesized that ratings obtained using the VALI scale included in this study might be less susceptible to the effects of auditory information because relative evaluation was not required for vibratory characteristics evaluated using the VALI, but was required for VAS ratings. However, results showed no main effect of auditory information on VAS ratings. Similarly, auditory information also did not strongly affect ratings using the VALI. In fact, there was a main effect of auditory information on only one VLS parameter measured using the VALI, non-vibrating portion-l. Across all levels of dysphonia severity, average VLS ratings for this parameter were decreased when auditory information was absent versus present, which differed from the non-significant increases in severity ratings for this parameter when measured using the VAS when auditory information was absent versus present (see Figures 2a and 3a). Because there was also a significant interaction with dysphonia severity for non-vibrating portion-l, interpretation of this finding will be discussed below. Additionally, no significant differences or trends were observed for non-vibrating portion-r or other VLS parameters for experimental VLS exams evaluated when auditory information was absent versus present. Similar to VAS ratings, there were no trends or significant differences for control VLS exams rated using the VALI when auditory information was absent versus present. Findings from ratings obtained using the VALI did not appear to be consistent with the presence of cognitive biases, as there were no systematic differences in VALI ratings observed when auditory cues were absent versus present.

The results of this study suggested that VALI ratings were not susceptible to cognitive biases. However, it is also important to consider how VLS parameters are measured using the VALI and/or how these ratings might be interpreted by clinicians. For example, the VALI includes both unipolar and bipolar scales with different ranges to measure some VLS parameters. Therefore, the direction of any effects of auditory information might vary by VLS parameter. This is in contrast to measurements made using a unipolar VAS scale in which increases in ratings reflect increased perceived severity, or increased deviance from normal. For example, increases in ratings for some VLS parameters (e.g., mucosal wave, amplitude of vibration), decreases for other parameters (e.g., non-vibrating portion) and both increases and decreases for other VLS parameters (e.g., phase closure) might all be associated with perceived decreases in deviance from normal. Because VALI ratings do not directly measure severity, they require additional interpretation. Lastly, some VLS parameters (e.g., amplitude of vibration, mucosal wave, non-vibrating portion) are measured using a similar range of scores, but are measured in different increments (e.g., 10%-20%). Consequently, some of these VLS parameters measured using the VALI might be more or less sensitive to change. Therefore, direct comparisons of ratings obtained using the 100 mm VAS and the VALI are not possible for many VLS parameters.

Together, our data showed that there was not a widespread effect of auditory information on VLS parameters using either rating tool, suggesting that auditory information contained within the VLS exam does not appear to affect average VLS ratings, regardless of the scale type. Inter-rater reliability for nearly all VLS parameters obtained using the VALI was similar to one prior study[28] and were improved compared to studies using a variety of rating scales and measures.[29,30] Inter-rater reliability for VALI ratings was also similar to those obtained using the

VAS scale in this study. In addition to the characteristics of rating scales used to obtain VLS ratings, dysphonia severity was predicted to potentially affect the direction and magnitude of differences in VLS ratings when auditory cues were absent versus present. Therefore, the interaction between dysphonia severity and auditory information was also examined.

Dysphonia Severity by Auditory Information

The effect of the interaction between dysphonia severity and auditory information on vibratory VLS parameters was evaluated because we predicted that the direction and/or the magnitude of the effect of auditory information might differ by level of dysphonia severity. For example, when speaker-specific information suggests the absence of a voice disorder, most studies have observed no effect or improvement in auditory-perceptual and visual-perceptual voice ratings, while speaker-specific information suggesting the presence of dysphonia severity typically results in increases in perceived severity.[10,34,35] Therefore, it was important to evaluate the interaction between dysphonia severity and auditory information to increase our ability to detect any effect of auditory information that depends on dysphonia severity. Additionally, previous studies have found that the magnitude of the effect of speaker-specific information is increased when a stimulus is mild in severity.[33,34] Because we presumed that speakers in this study with mild dysphonia might also have many VLS parameters that were rated as mild in severity, we hypothesized that the magnitude of the effect might be greatest for these speakers.

For VLS exams rated using the VAS, there were no significant interactions between auditory information and dysphonia severity for any VLS parameter. Using the VALI to obtain VLS ratings, there was only one statistically significant interaction between auditory cues and dysphonia severity, non-vibrating portion-l. Although some prior studies[33,34] have identified

greater effects of speaker-specific information on auditory-perceptual voice ratings and visual-perceptual ratings of VLS exams in the mild to moderate severity range, the effect of auditory cues did not appear to be greatest for speakers with mild dysphonia in this study. Instead, evaluation of the only significant interaction for non-vibrating portion-l showed that differences in VALI ratings increased with increasing dysphonia severity, resulting in statistically significant differences in ratings only when auditory information suggested moderate-severe dysphonia (see figure 4).

We also expected that there would be significant differences in average VLS ratings for speakers with different levels of dysphonia severity in this specific patient population (e.g., control and structural abnormalities without significant supraglottic hyperfunction). And, consistent with this hypothesis, we did find a main effect of dysphonia severity on most VLS parameters measured using the VAS and some VLS parameters evaluated using the VALI. This is important because as previously discussed, prior studies in broader fields of medicine[36,38] and auditory-perceptual voice assessment[35] found that the magnitude of the effect of speaker-specific information on perceptual rating tasks is reduced when speaker-specific information is accurate, or consistent with the stimulus, compared to inconsistent with the stimulus. However, our data suggests that increases in perceived severity of most VLS vibratory characteristics appeared to be consistent with increases in dysphonia severity, suggesting that the "accuracy" of the speaker-specific information provided by these auditory cues was adequately controlled in this study. Consequently, the lack of a significant effect of these "accurate" auditory cues on most VLS parameters was not surprising.

In this study, the non-significant differences in VLS ratings for most VLS parameters when auditory information was absent versus present have implications for future study designs.

Based on our results, we suggest that future studies should retain auditory information to increase the external validity of findings when the research purpose is not to compare high speed video exams with VLS. These recommendations are also based on the fact that any non-significant average differences in VLS ratings across auditory conditions in this study are unlikely to be clinically meaningful (see figures 2a and 2b). One caveat is that in addition to dysphonia severity, auditory information might provide clinicians with additional auditory cues (e.g., pitch, loudness, voice quality, etc.) in research and clinical settings. Although we controlled for fundamental frequency and intensity during the VLS selection process in this study, some auditory cues are important for making VLS ratings (e.g., ratings made at a comfortable pitch and loudness). Thus, in addition to dysphonia severity, other auditory cues should be considered when designing clinical and research protocols that include VLS exams with the auditory information retained. These factors need to be considered to maintain adequate rater reliability.

Clinical Decisions-Diagnostic Coding and Treatment

Although it is unlikely that the small observed average differences in vibratory VLS ratings in this study were clinically meaningful, it is possible that the auditory information provided in the VLS exams might have had a direct effect on clinical decisions about diagnosis or treatment. As outlined in figure 1, it was possible that auditory information could directly affect these clinical decisions, independent of the effects on VLS ratings. Therefore, this study was also designed to examine any possible effects of auditory information on an individual clinician's judgments about diagnostic coding or treatment recommendations.

Overall, there was a high level of agreement in both diagnostic codes and treatment recommendations in the absence and presence of auditory information, with disagreements in diagnostic codes and treatment recommendations occurring for approximately one-third of these observations. When disagreements in diagnostic coding and treatment recommendations were observed, a few trends were identified. For example, when disagreements about treatment recommendations were observed for experimental exams, a higher level of care (e.g., behavioral intervention versus surveillance) was more commonly recommended when auditory information was absent versus present. Overall, VLS severity ratings measured using the VAS scale were higher for experimental VLS exams when auditory cues were absent versus present. In other words, when disagreements were observed, average VLS parameters were rated as more severe and this was consistent with the increased frequency of individual clinicians recommending higher levels of care for experimental VLS exams when auditory information was absent.

When disagreements were observed for control VLS exams, VLS exams were commonly coded as within normal limits. In these cases, lower levels of care were more commonly recommended when auditory information was absent versus present. But, both increases and decreases in VLS severity ratings obtained using the VAS for the control exams were observed when auditory cues were absent versus present. As previously discussed, additional clinician interpretation is required to examine these relationships between VLS ratings obtained using the VALI and clinical decisions because these VLS ratings do not reflect a clinicians' perception of severity, or deviance from normal for many VLS parameters. Lastly, clinicians did not demonstrate any difference in average confidence ratings for diagnostic coding or treatment recommendations when auditory cues were present versus absent, suggesting that the presence of auditory cues did not appear to increase clinicians' confidence in making these types of clinical

judgments while performing the rating tasks. However, exit survey results indicated that clinicians, on average, anticipated that auditory cues were more likely to have influenced treatment recommendations compared to diagnostic coding or VLS ratings.

Although observations about diagnostic coding and treatment recommendations are interesting, further investigation is needed to determine the overall effect of auditory information suggesting different levels of dysphonia severity on diagnostic coding and treatment recommendations. This is important to consider because clinical diagnosis and levels of care are commonly used to validate clinical assessment tools, including those that rely on visual-perceptual judgments.[36,38,65] The presence of auditory information that could potentially affect the strength of the relationships between visual-perceptual VLS ratings and these types of clinical judgments should be considered in future laboratory study designs. However, findings related to diagnosis and treatment from this study and from other laboratory studies have limited external validity because clinicians were blinded to potentially important information (cognitive status, occupational voice demands, social circumstances, etc.) that might be used to make decisions about clinical diagnosis or treatment in a clinical setting.[66,67]

Limitations and Future Directions

Overall, the effect of auditory information on VLS ratings in this study was reduced in comparison to prior studies investigating other types of speaker-specific information, such as risk and protective factors, referring medical laryngeal diagnoses, and case histories suggesting specific voice disorder etiologies, on auditory- and visual-perceptual VLS ratings.[10,33–35] The observed effect of auditory information on average VLS ratings for experimental VLS exams measured using a 100 mm VAS was consistent with bias, but these differences were not

statistically significant and were unlikely to be clinically meaningful. While retaining auditory information during clinical protocols and future study designs might be useful for optimizing external validity without significantly reducing the internal validity of studies that include VLS, there are several additional considerations for interpreting the results of this study.

Inter-rater reliability for most VLS parameters measured with or without auditory information and measured using the VAS and the VALI was acceptable and even superior to inter-rater reliability previously reported for many VLS parameters,[28–30] reflecting adequate stability in most perceptual measures using these rating methodologies. However, the effect of speaker-specific information on VLS ratings might vary with the stability of the perceptual parameter being rated using other methods. Despite high levels of inter-rater reliability observed in this study, intra-rater reliability was highly variable for many VLS ratings obtained using the VAS and the VALI. Because clinicians rated all videos in a single rating session to control for variability in the rating tasks, measures of intra-rater reliability are based on repeated ratings of only 2 of 16 VLS exams so that rater burden and fatigue were reduced. Repeating only two VLS exams likely negatively impacted measures of intra-rater reliability. This is an important consideration for future study designs. In addition to the overall stability of perceptual measures, the order and amount of speaker-specific information that is presented should also be considered.

In this study, the auditory information was provided to clinicians simultaneously with the VLS, whereas previous studies[10,33–35] and in clinical practice, speaker-specific information is present prior to evaluating VLS exams. The effect of the order that speaker-specific information is provided has not been investigated. Secondly, we controlled for the amount of auditory information that was present during the VLS exams by including only comfortable-pitched and high-pitched sustained vowel productions obtained from a specific patient population. However,

the amount of auditory information was very limited in this study compared to VLS exams obtained in clinical studies and/or those that are obtained using other instrumentation. For example, a flexible versus rigid endoscope is often used to obtain VLS exams because additional speech and voice tasks can be evaluated. The effect of those auditory cues might vary from the effect of the limited auditory cues evaluated in this study.

Previous studies have also observed differences in the direction and magnitude of the effects of different types of speaker-specific information when it is consistent versus inconsistent with the stimulus. This study attempted to control for the consistency of the auditory information with the severity of the VLS parameter being rated and/or with overall VLS exam severity using auditory-perceptual voice ratings of the audio signal extracted from the VLS exam. However, because the microphone to equipment distance was not controlled during data acquisition, noise levels from the equipment was likely variable between exams. This could have affected the auditory-perceptual dysphonia severity ratings obtained from experienced SLPs and reduced our level of control over the accuracy of the auditory cues. However, the significant main effect of dysphonia severity on many VLS parameters measured using either a 100 mm VAS or the VALI and overall severity of laryngeal function using a 100 mm VAS suggested that measures of dysphonia severity were useful for categorizing VLS severity in this patient population.

Although the consistency of speaker-specific information with the VLS exam appeared well controlled in this study, some patient populations (e.g., spasmodic dysphonia, functional dysphonia) or individual speakers might exhibit voice characteristics that are highly variable or task specific and might therefore be inconsistent with the VLS stimulus. For example, a speaker might present with severely abnormal auditory-perceptual voice quality during perceptual voice assessment prior to obtaining a VLS exam, but demonstrate improved voice quality during the

VLS exam or during certain speech or voice tasks obtained during VLS. Future studies would be needed to determine if the effect of auditory-perceptual information on VLS ratings and clinical judgments might depend on factors such as the amount, type, consistency of the auditory perceptual information with the VLS stimulus, and/or order of presentation.

As previously discussed, the observed differences in ratings of VLS parameters as a function of auditory information were unlikely to be clinically meaningful when measured using the VAS or the VALI. In addition, the diagnostic codes and treatment recommendations provided by clinicians upon evaluation of VLS exams with and without auditory information were commonly in agreement. However, some observations suggest that the presence of auditory cues might directly affect diagnostic coding and treatment recommendations in a laboratory setting, and should be further examined. Lastly, only experienced SLP clinicians were included in this study. Therefore, the results cannot be generalized to clinicians with different professional backgrounds or to novice clinicians. Because VLS is considered the most important instrumental clinical voice assessment tool for evaluating individuals with dysphonia, it is hoped that future investigations will be useful for increasing the validity of these measures performed in clinical and laboratory settings.

References

1. ASHA Practice Portal. https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942600&section=Overview.
2. Emerick LL, Van Riper C. *Speech Correction : An Introduction to Speech Pathology and Audiology*. 8th ed. Englewood Cliffs, N.J.: Englewood Cliffs, N.J. : Prentice Hall; 1990.
3. Cohen SM, Kim J, Roy N, Courey M. Delayed otolaryngology referral for voice disorders increases health care costs. *Am J Med*. 2015;128(4):426.e11-18.
4. Rosen CA, Lombard LE, Murry T. Acoustic, Aerodynamic, and Videostroboscopic Features of Bilateral Vocal Fold Lesions. *Annals of Otology, Rhinology & Laryngology*. 2000;109(9):823-828.
5. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am J Speech Lang Pathol*. 2018;27(3):887-905.
6. Behrman, Alison. "Common Practices of Voice Therapists in the Evaluation of Patients." Journal of Voice, vol. 19, no. 3, 2005, pp. 454–469.
7. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-based clinical voice assessment: a systematic review. *Am J Speech Lang Pathol*. 2013;22(2):212-226.
8. Paul BC, Chen S, Sridharan S, Fang Y, Amin MR, Branski RC. Diagnostic accuracy of history, laryngoscopy, and stroboscopy. *Laryngoscope*. 2013;123(1):215-219.
9. Cohen SM, Kim J, Roy N, Wilk A, Thomas S, Courey M. Change in diagnosis and treatment following specialty voice evaluation: A national database analysis. *Laryngoscope*. 2015;125(7):1660-1666.
10. Sauder C, Nevdahl M, Kapsner-Smith M, Merati A, Eadie T. Does the accuracy of case history affect interpretation of videolaryngostroboscopic exams? *The Laryngoscope*. 2020;130(3):718-725.
11. Bonilha HS, Focht KL, Martin-Harris B. Rater methodology for stroboscopy: a systematic review. *J Voice*. 2015;29(1):101-108.
12. Simpson C. Simpson, C., Linda May, Jill Green, Robert Eller, and Carlayne Jackson. "Vibratory Asymmetry in Mobile Vocal Folds: Is It Predictive of Vocal Fold Paresis?" The Annals of Otology, Rhinology & Laryngology 120, no. 4 (2011): 239-42.
13. El-Demerdash. El-Demerdash, A., Fawaz, S. A., Sabri, S. M., Sweed, Ahmed, and Rabie, H. "Sensitivity and Specificity of Stroboscopy in Preoperative Differentiation of Dysplasia from Early Invasive Glottic Carcinoma." European Archives of Oto-Rhino-Laryngology 272, no. 5 (2015): 1189.
14. Sataloff R. Sataloff, R., Mandel, S., Mann, E., & Ludlow, C. (2004). Practice Parameter: Laryngeal Electromyography (An Evidence-Based Review). Otolaryngology–Head and Neck Surgery, 130(6), 770-779.
15. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-Based Clinical Voice Assessment: A Systematic Review. 2013;22(2):212-226.
16. Estes C. Estes, Christine, et al. "Laryngoscopic and Stroboscopic Signs in the Diagnosis of Vocal Fold Paresis." The Laryngoscope, vol. 127, no. 9, 2017, pp. 2100–2105.
17. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *J Speech Hear Res*. 1993;36(1):21.

18. Jones JW, Perryman M, Judge P, et al. Resident Education in Laryngeal Stroboscopy and Perceptual Voice Evaluation: An Assessment. *Journal of Voice*. 2018.
19. Ferrer C. Ferrer, Carlos A., et al. "Collinearity and Sample Coverage Issues in the Objective Measurement of Vocal Quality: The Case of Roughness and Breathiness." Journal of Speech, Language, and Hearing Research, vol. 61, no. 1, 2018, pp. 1–24.
20. Oates J. Auditory-Perceptual Evaluation of Disordered Voice Quality. *Folia Phoniatr Logop*. 2009;61(1):49-56.
21. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America*. 2000;108(4):1867-1876.
22. Poburka B. A Multi-Media, Computer-Based Method for Stroboscopy Rating Training." Journal of Voice : Official Journal of the Voice Foundation., vol. 12, no. 4, 1998, pp. 513–526.
23. Chan KMK, Yiu EM-L. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45(1):111-126.
24. Eadie TL, Kapsner-Smith M. The Effect of Listener Experience and Anchors on Judgments of Dysphonia. *J Speech Lang Hear Res*. 2011;54(2):430-447.
25. Lowell SY, Kelley RT, Busekroos L, et al. The effect of anchors on reliability of endoscopic tremor ratings: Anchors for Endoscopic Tremor Ratings. *The Laryngoscope*. 2017;127(2):411-416.
26. Bonilha, H. S., Desjardins, M. L., Garand, K., & Martin-Harris, B. (2017). Parameters and Scales Used to Assess and Report Findings From Stroboscopy: A Systematic Review. Journal of Voice, .
27. DeCastellarnau A. A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity*. 2018;52(4):1523-1559.
28. Poburka BJ, Patel RR, Bless DM. Voice-Vibratory Assessment With Laryngeal Imaging (VALI) Form: Reliability of Rating Stroboscopy and High-speed Videoendoscopy. *J Voice*. 2017;31(4):513.e1-513.e14.
29. Rosen CA. Stroboscopy as a research instrument: development of a perceptual evaluation tool. *Laryngoscope*. 2005;115(3):423-428.
30. Nawka T, Konerding U. The inter-rater reliability of stroboscopy evaluations. *J Voice*. 2012;26(6):812.e1-10.
31. Patel R, Dailey S, Bless D. Comparison of High-Speed Digital Imaging with Stroboscopy for Laryngeal Imaging of Glottal Disorders. *Ann Otol Rhinol Laryngol*. 2008;117(6):413-424.
32. *Hirano M, Bless DM. (1993) Videostroboscopic Examination of the Larynx. San Diego, CA: Singular Publishing Group, Inc.*
33. Teitler N. Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. *J Voice*. 1995;9(1):95-105.
34. Eadie T, Sroka A, Wright DR, Merati A. Does Knowledge of Medical Diagnosis Bias Auditory-Perceptual Judgments of Dysphonia? *Journal of Voice*. 2011;25(4):420-429.
35. Sauder C, Eadie T. Does the Accuracy of Medical Diagnoses Affect Novice Listeners' Auditory-Perceptual Judgments of Dysphonia Severity? *Journal of Voice*. 2020;34(2):197-207.
36. Leblanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. *Acad Med*. 2002;77(10 Suppl):S67-69.

37. Bytzer P. Information Bias in Endoscopic Assessment. *American Journal of Gastroenterology*. 2007;102(8):1585-1587.
38. Hatala R, Norman GR, Brooks LR. Impact of a clinical scenario on accuracy of electrocardiogram interpretation. *J Gen Intern Med*. 1999;14(2):126-129.
39. Shikino K, Ikusaka M, Ohira Y, et al. Influence of predicting the diagnosis from history on the accuracy of physical examination. *Adv Med Educ Pract*. 2015;6:143-148.
40. Hatala RM, Brooks LR, Norman GR. Practice makes perfect: the critical role of mixed practice in the acquisition of ECG interpretation skills. *Adv Health Sci Educ Theory Pract*. 2003;8(1):17-26.
41. Kahneman D. Kahneman, D., Slovic, P., & Tversky A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.
42. Hammond KR. How Convergence of Research Paradigms Can Improve Research on Diagnostic Judgment. *Medical Decision Making*. 1996;16(3):281-287.
43. Yiu, E., Lau, V., Ma, E., Chan, K., & Barrett, E. (2014). Reliability of laryngostroboscopic evaluation on lesion size and glottal configuration: A revisit. The Laryngoscope, 124(7), 1638-44.
44. Roy N, Barton ME, Smith ME, Dromey C, Merrill RM, Sauder C. An in vivo model of external superior laryngeal nerve paralysis. *Laryngoscope*. 2009;119(5):1017–1032.
45. Isseroff TF, Parasher AK, Richards A, Sivak M, Woo P. Inter-rater Reliability in Analysis of Laryngoscopic Features for Unilateral Vocal Fold Paresis. *J Voice*. October 2015.
46. Ludlow C. Ludlow, C., Domangue, R., Sharma, D., Jinnah, H., Perlmutter, J., Berke, G., . . . Stebbins, G. (2018). Consensus-Based Attributes for Identifying Patients With Spasmodic Dysphonia and Other Voice Disorders. JAMA Otolaryngology-- Head & Neck Surgery, JAMA otolaryngology-- head & neck surgery, 21 June 2018.
47. Martin-Harris B, Brodsky MB, Michel Y, et al. MBS Measurement Tool for Swallow Impairment—MBSImp: Establishing a Standard. *Dysphagia*. 2008;23(4):392-405.
48. Sulica L. Vocal Fold Paresis: An Evolving Clinical Concept. *Curr Otorhinolaryngol Rep*. 2013;1(3):158-162.
49. Smith M, Higgs J, Ellis E. Characteristics and processes of physiotherapy clinical decision making: a study of acute care cardiorespiratory physiotherapy. *Physiother Res Int*. 2008;13(4):209-222.
50. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J*. 1987;66(7):289-296.
51. KayPENTAX. Instruction manual: Stroboscopy systems and components. Montvale, NJ: 2008.
52. Makki F. Makki, F., Hilal, A., Fung, E., Hart, R., Taylor, S., & Brown, T. (2013). Accuracy of flexible versus rigid laryngoscopic photo-documentation in the diagnosis of early glottic cancer. The Journal of Laryngology and Otology, 127(9), 890-6.
53. Rzepakowska A, Sielska-Badurek E, Osuch-Wojcikiewicz E, Sobol M, Niemczyk K. The predictive value of videostroboscopy in the assessment of premalignant lesions and early glottis cancers. *Otolaryngol Pol*. 2017;71(4):14-18.
54. Uloza V. Uloza, Vegienė, & Šaferis. (2013). Correlation Between the Basic Video Laryngostroboscopic Parameters and Multidimensional Voice Measurements. Journal of Voice, 27(6), 744-752.

55. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18(2):124-132.
56. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20(1):14-22.
57. Heller A, Tanner K, Roy N, et al. Voice, Speech, and Laryngeal Features of Primary Sjögren's Syndrome. *Annals of Otology, Rhinology & Laryngology*. 2014;123(11):778-785.
58. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21(5):576-590.
59. Allensworth. Allensworth, O'Dell, Ziegler, Bryans, Flint, & Schindler. (2017). Treatment Outcomes of Bilateral Medialization Thyroplasty for Presbylaryngis. Journal of Voice, .
60. *IBM Corp. Released 2017. IBM SPSS Statistics for Mac, Version 26.0. Armonk, NY: IBM Corp.*
61. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979;86(2):420-428.
62. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23-34.
63. Portney L, Watkins M. Construct validity. Foundations of Clinical Research: Applications to Practice. *Prentice Hall Health New Jersey, USA*. 2000:87-91.
64. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
65. Martin-Harris B, Brodsky MB, Michel Y, et al. MBS Measurement Tool for Swallow Impairment—MBSImp: Establishing a Standard. *Dysphagia*. 2008;23(4):392-405.
66. Bachmann L. Bachmann, Lucas M., Muhleisen, Andrea, Bock, Annekatrin, Ter Riet, Gerben, Held, Ulrike, and Kessels, Alfons GH. "Vignette Studies of Medical Choice and Judgement to Study Caregivers'' Medical Decision Behaviour: Systematic Review.(Research Article)." BMC Medical Research Methodology 8, no. 50 (2008): 50."
67. Gigerenzer G. Gigerenzer, Gerd. "Does the Environment Have the Same Structure as Bayes'' Theorem?" Behavioral and Brain Sciences, vol. 14, no. 03, 1991, pp. 495–496."

Appendix A

Videolaryngostroboscopic Rating Form

Instructions: Please take as much time as you need to become familiar with the following definitions and the way that we would like you to rate the following laryngoscopic and stroboscopic parameters prior to viewing the videos. The videos were obtained from adult males and females with and without voice symptoms. You can advance frame by frame, pause, and resume playing the videos as often as needed. Each video will include a recording of a speaker sustaining phonation at a self-perceived comfortable pitch and loudness level. Each sustained phonation will be repeated 3 times. Then, a self-perceived high pitched sustained phonation is repeated 3 times. **Please make all ratings during productions at comfortable pitch and loudness (first 3 productions), unless otherwise indicated on rating form.** You can play the video in its entirety a total of 3 times. Although you cannot view the video more than three times, you can take as much time as needed after each video viewing to complete the rating form, briefly provide treatment recommendations, and provide the ICD-10 code(s) that you would use to bill a 3rd party payer for voice assessment. Begin playing the first video when you are ready.

Gross Laryngeal Structure and Motion

1A.

Free Edge Contour					
Definition:	Smoothness &/or straightness of free edge.				
Rating:	Rate right & left VFs separately during abduction. Write in one rating per vocal fold.				
					
Normal	Convex	Concave	Irregular	Rough	
Right:___			Left:___		

1B.

Draw a tic mark that reflects the degree of abnormality of the **free edge contour**

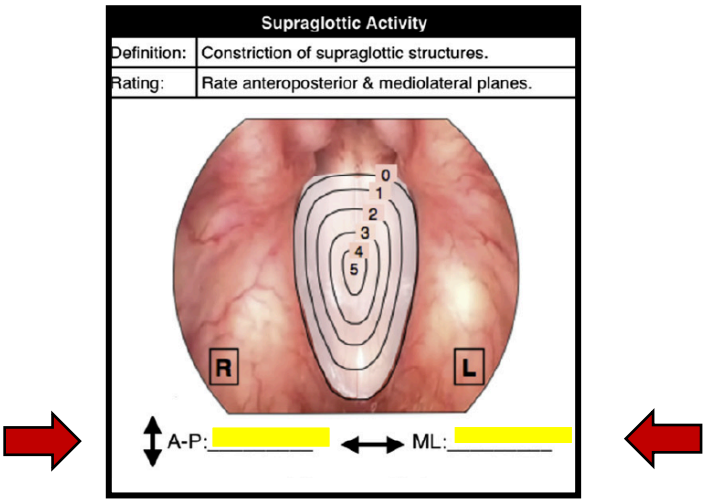
Right

Normal — Profoundly Abnormal

Left

Normal — Profoundly Abnormal

2A.



2B.

Draw a tic mark that reflects the degree of abnormality for **supraglottic activity**

A-P

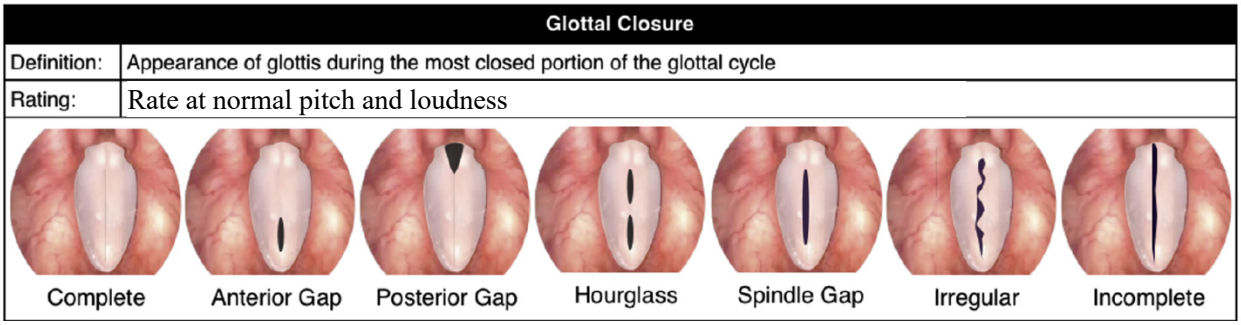
Normal — Profoundly Abnormal

M-L

Normal — Profoundly Abnormal

VIBRATORY CHARACTERISTICS

3A.



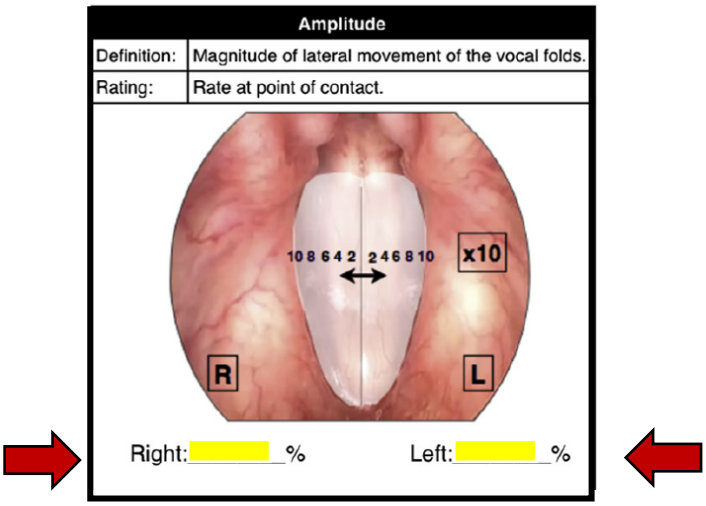
Glottal Closure_____(choose 1)

____ **3B.**

Draw a tick mark that reflects the degree of abnormality for **glottal closure**

Normal — Profoundly Abnormal

____ **4A.**



____ **4B.**

Draw a tick mark that reflects the degree of abnormality for **amplitude**

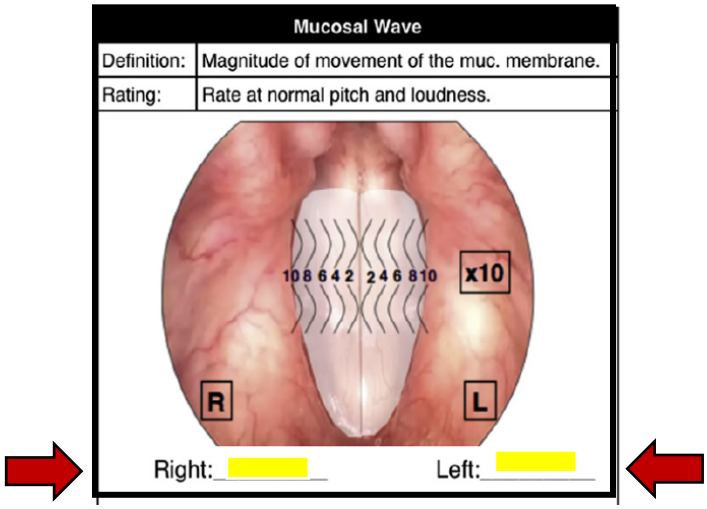
Right

Normal — Profoundly Abnormal

Left

Normal — Profoundly Abnormal

____ **5A.**



____ **5B.** Draw a tick mark that reflects the degree of abnormality for **mucosal wave**

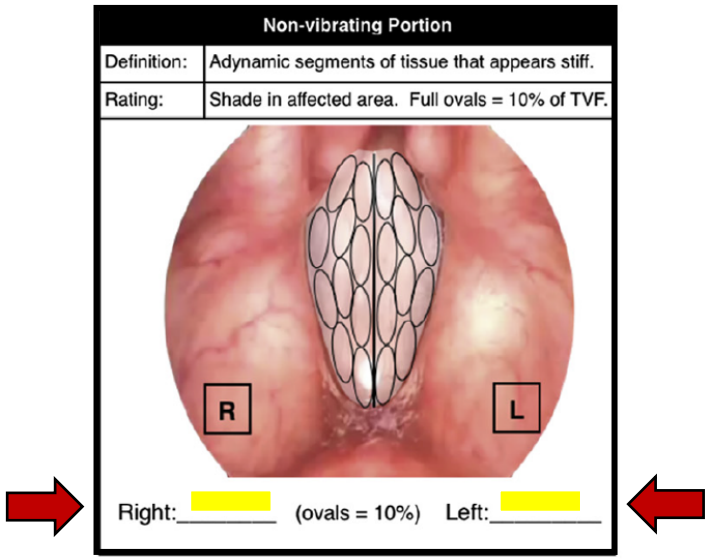
Right

Normal — Profoundly Abnormal

Left

Normal — Profoundly Abnormal

____ **6A.**



____ **6B.** Draw a tic mark that reflects the degree of abnormality for **non-vibrating portion**

Right

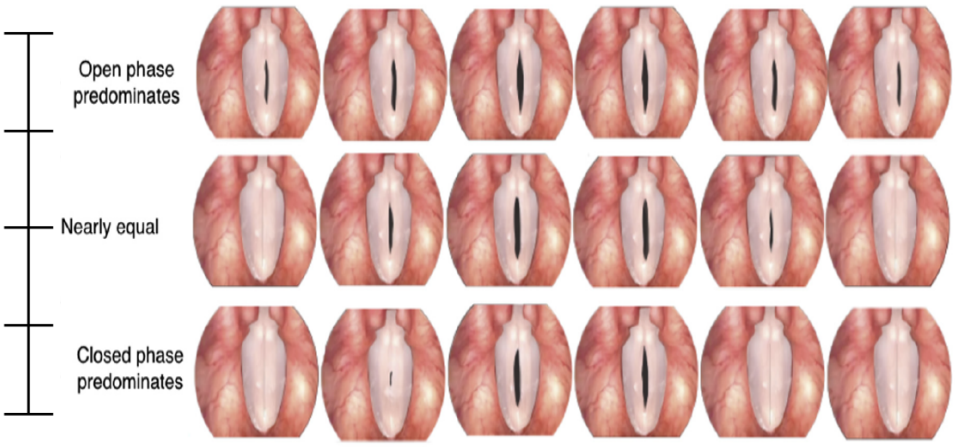
Normal — Profoundly Abnormal

Left

Normal — Profoundly Abnormal

7A.

Phase Closure	
Definition:	The relative durations of appearance of consecutive glottal cycles.
Rating:	Circle a hash mark on the continuum below. Rate at point of contact.

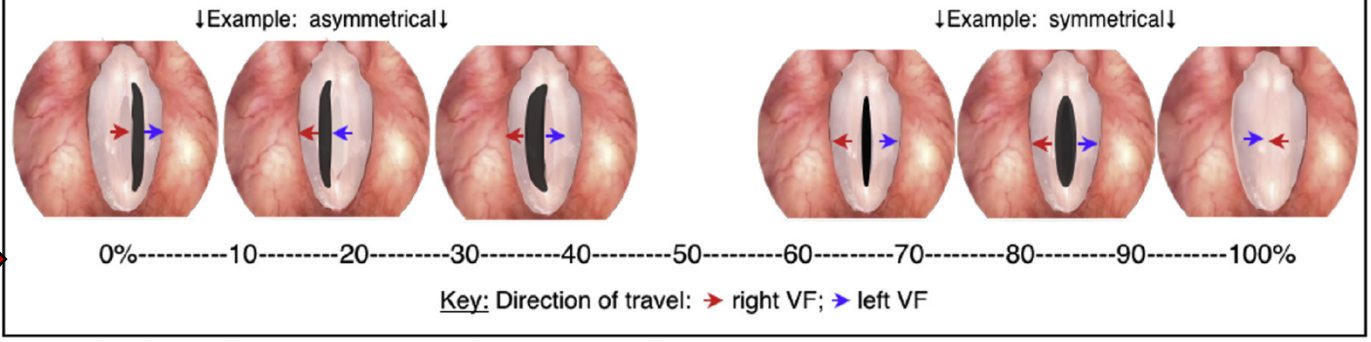


7B. Draw a tic mark that represents the degree of abnormality for **phase closure**

Normal — Profoundly Abnormal

8A.

Phase Symmetry	
Definition:	The degree to which the VFs move as mirror-opposite images of each other (180° phase difference).
Rating:	Select the % of exam time that vibration is symmetrical.



8B.

Draw a tic mark that represents the degree of abnormality for **phase symmetry**

Normal — Profoundly Abnormal

____ **9A.**

Regularity	
Definition:	Consistency of cycles.
Rating:	Circle % of time that vibration is regular

never reg. 0% -10 - 20 - 30 - 40 - 50 - 60 -70 - 80 - 90 -100% always reg.

poor tracking; unstable/blurred imaging — good tracking; stable imaging

____ **9B.** Draw a tic mark that represents the degree of abnormality for **phase regularity**

Normal — Profoundly Abnormal

____ **10.** Draw a tic mark that reflects **<u>overall impression</u>** of **laryngeal function**

Normal — Profoundly Abnormal

____ **11.** Was stroboscopic tracking ***<u>adequate</u>*** for rating (circle)?

YES NO

Please mark your response to the following questions based on the information provided by the videolaryngostroboscopic exam.

12. Of the options provided, check the **best primary treatment recommendation** (choose only 1) based on the information provided:

___1) **No treatment**

___2) **Surveillance/Monitor**

___3) **Vocal hygiene only**

___4) **Voice rest/conservation**

___5) **Behavioral voice therapy** (e.g. vocal function exercises, resonant voice, etc.)

___6) **Medication** (e.g. steroids, reflux, etc.)

___7) **Surgical intervention** (e.g. excision, medialization, etc.)

13. Please rate your **level of confidence** in the **treatment recommendation** above (circle 1).

Really Confident	Confident	Not sure	Not confident	Not really confident

14. Select the most specific primary medical laryngeal diagnosis (see list of medical diagnoses and their associated ICD-10 codes) that you would use for billing purposes or circle within normal limits (WNL). If the primary medical diagnosis differs by vocal fold, please provide a second medical laryngeal diagnosis or select WNL. (For example: J38.3 **right** cyst of the vocal cords/larynx and J38.7 **left** vocal fold hemorrhage)

Medical laryngeal diagnosis _____with associated ICD-10 _____ **OR** WNL

_____Right vocal fold

_____Left vocal fold

14b. ONLY if the medical laryngeal diagnosis differs between the right and left vocal fold, please provide a second medical laryngeal diagnosis or circle WNL.

Medical laryngeal diagnosis _____ with associated ICD-10 _____ **OR** WNL

_____ Right vocal fold

_____ Left vocal fold

15. Please rate your **level of confidence** in the medical diagnosis and associated ICD-10 code(s) selected (circle 1).

Really Confident	Confident	Not sure	Not confident	Not really confident

Appendix B

C32.9 **Cancer of the larynx/vocal fold**

D14.1 **Benign neoplasm of larynx** (includes papillomatosis)

J04.0 **Acute laryngitis**

J37.0 **Chronic laryngitis** (Catarrhal laryngitis, Hypertrophic laryngitis, Sicca laryngitis)

J38.0 **Paralysis/Paresis** of vocal cords and larynx

J38.1 **Polyp** of vocal cord and larynx (Excludes1: adenomatous polyps (D14.1))

J38.2 **Nodules** of vocal cords

J38.3 **Abscess** of vocal cords

J38.3 **Cyst** of the vocal cords/larynx

J38.3 **Cellulitis** of vocal cords

J38.3 **Granuloma** of vocal cords

J38.3 **Leukokeratosis** or **leukoplakia** of vocal cords

J38.3 **Sulcus vocalis** or **scar** of vocal fold

J38.4 **Edema of larynx** (Edema (of) glottis, Subglottic edema, Supraglottic edema)

J38.7 **Atrophy of the vocal fold**

J38.7 **Vocal fold hemorrhage**

J38.7 **Vocal fold varix**

K21.9 **Gastro-esophageal reflux disease without esophagitis** (laryngopharyngeal reflux, reflux laryngitis)

R49.0 **Dysphonia/hoarseness**

Chapter 5

General Discussion and Summary of Findings

1. General Discussion

The broad purpose of this dissertation was to investigate one potential source of variability, the effect of exposure to different types of information about a speaker, on auditory-perceptual and visual-perceptual videolaryngostroboscopy (VLS) rating tasks and other clinical judgments about voice diagnosis and/or treatment. Exposure to speaker-specific information differs between a laboratory and a clinical setting, but the effect of exposure to different types of speaker-specific information on auditory- and visual-perceptual rating tasks has not been systematically investigated. These perceptual voice assessment measures are considered to be the most common and important clinical assessment tools used to determine the presence, severity, and/or nature of a voice disorder, assess treatment outcomes, and determine prognosis for improvement with treatment(s) in a clinical setting.[1,2] Therefore, understanding the sources of variability in these perceptual rating tasks is important for improving the validity of perceptual measures performed in a laboratory and/or clinical setting. Investigating the effect of speaker-specific information on perceptual voice rating tasks is also important for evaluating the external validity of laboratory studies in which clinicians are not typically exposed to speaker-specific information during perceptual voice rating tasks.

In this dissertation, Chapter 1 described an existing conceptual model of how an acoustic signal is mapped to auditory-perceptual voice assessment and how this model could be adapted and extended to provide a framework for explaining and testing factors that affect the validity of VLS (refer to figures 2 and 3 in Chapter 1).[3] In Chapters 2, 3, and 4, we used this conceptual framework to design studies that could control/account for many known and/or potential sources of variability in mapping a stimulus onto auditory-perceptual or visual-perceptual VLS ratings,

such as those associated with the quality of the stimuli, the rater (e.g., profession, experience, etc.) and other rating task factors (e.g., stimulus order, types of stimuli, rating tools, etc.). At the same time, the research designs used in Chapters 2, 3, and 4, allowed us to focus our investigations on the effect of speaker-specific information on perceptual voice assessment measures.[3]

In Chapter 2, we investigated the effect of a referring medical laryngeal diagnoses (e.g., benign lesion(s), paralysis/paresis, control/other) on associated auditory-perceptual voice ratings of roughness and breathiness. We also examined whether the effect of these referring medical laryngeal diagnoses was dependent on their accuracy. In Chapter 3, we investigated the effect of case histories suggesting a specific voice disorder etiology (e.g., paresis, reflux, hemorrhage) on visual-perceptual VLS ratings of associated visual-perceptual parameters (e.g., vocal fold mobility, posterior laryngeal appearance, mucosal wave). However, the VLS exams evaluated in Chapter 3 contained an audio signal that might have affected the results. Consequently, in Chapter 4, we investigated the effect of retaining the audio signal in VLS exams on visual-perceptual vibratory VLS ratings. We also examined whether the effect might depend on the level of dysphonia severity or the type of rating scale used for measurement. Lastly, Chapters 3 and 4 explored the effect of speaker-specific information on judgments about clinical diagnosis or treatment.

2. Summary findings

In Chapter 2, we investigated the effect of knowing a referring medical laryngeal diagnosis on clinicians' auditory-perceptual ratings of voice quality. This was an important factor to consider because medical diagnoses are often known prior to voice evaluations, as

speech-language pathologist often receive referrals from physicians. One other important consideration was related to the accuracy of these referring diagnoses; one previous study[7] reported that half of all patients referred to a specialty voice clinic from an outpatient otolaryngology practice had a change in laryngeal diagnosis following specialized voice assessment.

In summary, we found a significant effect of referring medical laryngeal diagnoses on 40 novice clinicians' ratings of breathiness and roughness in 30 speech samples. The direction of the effect differed by the type of diagnostic label provided to listeners. For example, we observed *increases* in average perceived breathiness and roughness when medical laryngeal diagnoses suggested paralysis/paresis or benign lesions. We also observed *decreases* in average perceived severity when medical laryngeal diagnoses either indicated normal larynx or a medical laryngeal diagnosis that was not specifically associated with these auditory-perceptual parameters (i.e., roughness, breathiness). Further, we found that the magnitude of the effect of the medical laryngeal diagnoses was dependent on the accuracy of the diagnostic label provided to clinicians: In comparison to a control condition when no diagnosis was present, differences in clinicians' perceived severity of breathiness and roughness were increased when the diagnostic label was inaccurate versus accurate

In Chapter 3, we extended our study of the effect of diagnostic information to visual-perceptual voice evaluation. Similar to the results reported in Chapter 2, we found an effect of case histories that strongly suggested a particular voice disorder etiology on 2 of 3 visual-perceptual VLS parameters uniquely associated with these voice disorder etiologies in 32 experienced otolaryngologist and speech-language pathologist voice clinicians. Specifically, clinicians were more likely to detect and rate a VLS parameter as more severe when a case

history suggested its presence. These abnormalities were rated as less severe or were undetected when the case history suggested no abnormality. These findings illustrated differences in the direction of the effect when speaker-specific information suggested the presence or absence of a particular voice disorder etiology. We also found that case histories influenced clinical impressions about voice disorder diagnosis and that they also affected treatment recommendations.

In Chapter 3, accurate case histories were associated with improved overall agreement in clinical impressions compared to inaccurate case histories. Furthermore, case history information significantly predicted recommendations for treatment above and beyond VLS alone. Specifically, when a case history contained no self-reported voice symptoms (control), clinicians were unlikely to recommend treatment even when laryngeal pathology existed. One limitation of the study in Chapter 3 was that the VLS exams contained an audio signal that might also have provided clinicians with unique speaker-specific information about the etiology or the severity of the speaker's voice disorder. This was an important factor to consider because clinicians often are exposed to a patient's voice quality prior to a voice evaluation, and/or they often make VLS judgments in the presence of auditory cues that are contained in the VLS audio signal.

Thus, in Chapter 4 we evaluated the effect of the VLS audio signal on 9 visual-perceptual VLS parameters rated by 8 experienced SLP voice clinicians using a repeated measures study design. Clinicians in this study made judgments of 12 experimental exams and 4 control exams with or without auditory information. The results showed that there was no effect of auditory cues suggesting different levels of dysphonia severity (none, mild, moderate-severe) on the average perceived severity of any of the VLS parameters rated using 100 mm visual analog scales (VAS). Similarly, there was no effect of auditory cues on ratings of overall perceived

severity of laryngeal function measured using a 100 mm (VAS). When these same 9 stroboscopic parameters were rated using the Voice-Vibratory Assessment with Laryngeal Imaging (VALI), there was a significant main effect of auditory information and an interaction with dysphonia severity for only 1 of 9 stroboscopic parameters (non-vibrating portion-left (l)) (refer to figure 4 in Chapter 4). There was also a high level of agreement in diagnostic billing codes and treatment recommendations when individual clinicians evaluated VLS exams with auditory information present versus absent. Although it appeared that the effect of the audio signal during VLS was reduced in Chapter 4 compared to other types of speaker-specific information (e.g., referring medical laryngeal diagnosis, case history), there are several important factors that need to be considered in order to interpret the results of studies included in this dissertation. These factors will be discussed in the following sections.

2.1 Type of speaker-specific information

In order to compare the effects of different types of speaker-specific information (e.g., referring medical laryngeal diagnosis, risk/protective factors, case histories, auditory information) on auditory- and visual-perceptual VLS ratings or other clinical judgments, it is important to consider what types of cues about a speaker this information might provide. For example, in Chapters 2 and 3, referring medical laryngeal diagnoses and case histories provided clinicians with information about the absence/presence and etiology of a speaker's voice disorder. Based on their underlying pathophysiology, the voice disorder etiologies suggested by speaker-specific information in Chapters 2 and 3 are commonly associated with specific auditory- and visual-perceptual features. For example, speakers with voice disorders secondary to vocal fold lesions might be expected to have increased roughness and reduced mucosal wave during auditory- and visual-perceptual VLS voice assessments, respectively. Therefore, we

hypothesized that these perceptual features would be more likely to be detected and/or perceived as more severe when speaker-specific information that suggested these types of voice disorder etiologies were present. Findings from Chapters 2 and 3 were largely consistent with this hypothesis. Additionally, this effect carried over to clinical decision-making with regards to treatment. Clinicians were more likely to recommend treatment when an experimental case history that suggested the presence of a voice disorder versus the absence of a voice disorder was presented to clinicians in Chapter 3. In other fields of medicine, these results are consistent with systematic deviations from a normative response, also known as heuristics and cognitive biases.[4] Heuristics and biases can ultimately help or hinder clinicians who seek outcomes that confirm their expectations during medical examinations.

In contrast to the information provided in Chapters 2 and 3, the auditory information provided to clinicians in Chapter 4 did not suggest a particular voice disorder etiology. Instead, auditory cues provided information about the presence/absence and severity of dysphonia. We predicted that when VLS exams were evaluated in the presence of auditory cues that suggested increased dysphonia severity, ratings for many VLS vibratory characteristics and overall laryngeal function would be perceived as more severe compared to VLS exams evaluated in the absence of the auditory cues. However, none of the VLS parameters or overall severity of laryngeal function ratings demonstrated significant differences in the presence or absence of the audio signal when they were measured using a VAS. Further, only 1 visual-perceptual parameter was affected by the presence of auditory cues suggesting dysphonia when this parameter was measured using the VALI. Thus, it appears that the magnitude of the effect of these auditory cues suggesting increased severity of dysphonia presented simultaneously with the VLS exams may be less powerful cues than medical diagnoses presented prior to VLS exams. However,

differences in study designs, the accuracy of the speaker-specific information, the order that speaker-specific information was presented, the stability of perceptual ratings, and other factors must also be taken into account. These factors will be discussed further below.

The effect of speaker-specific information intended to serve as a control varied between and within studies in Chapters 2, 3, and 4. For example, there was no effect of *accurate* control medical laryngeal diagnoses that were either associated with normal laryngeal function and/or were not specifically associated with the auditory-perceptual outcome variables of interest in Chapter 2. This finding was consistent with one prior study of the effect of accurate medical laryngeal diagnoses on auditory-perceptual ratings obtained from both experienced and novice SLP clinicians.[5] However, there were decreases in perceived average severity of auditory-perceptual ratings with exposure to *inaccurate* control medical laryngeal diagnoses compared to no diagnosis in Chapter 2. Therefore, the direction and magnitude of the effect of the control speaker-specific information was dependent on its accuracy, which will be discussed further below.

In Chapter 3, a control case history resulted in decreases in perceived severity of visual-perceptual VLS ratings for most outcome variables. In other words, clinicians perceived VLS ratings as "more normal" when they performed their judgments in the presence of control case history information. This finding was consistent with a prior study by Teitler.[6] In that study, decreases in perceived severity of vibratory VLS parameters were observed when case histories suggested protective factors (e.g., adequate hydration, no smoking history, minimal voice demands) versus risk factors (e.g., poor vocal hygiene, smoking history, etc.) for voice disorders.[6] Clinicians were also less likely to recommend treatment when this control case history was presented for all VLS exams in Chapter 3.

Similarly, in Chapter 4, there was no effect of auditory cues that suggested the absence of dysphonia on VLS ratings. The effect of control speaker-specific information is important to consider because differences in the direction and magnitude of the effect of control versus experimental speaker-specific information can affect the *overall* magnitude of the effect. For example, when control speaker-specific information results in decreases in perceived severity and experimental speaker-specific information results in increases in perceived severity, the overall magnitude of this effect will appear reduced in comparison to no speaker-specific information unless it is accounted for in the study design. For this reason, these factors should be considered in future study designs and when interpreting the results of these studies. As previously mentioned, there is evidence that the overall magnitude of the effect of speaker-specific information depends on the accuracy, or the consistency of speaker-specific information with the stimulus being rated. Consequently, results from Chapters 2, 3, and 4 should also be considered in the context of the accuracy of this information.

2.2 Accuracy of Speaker-specific Information

Findings herein suggest that not only the *type* of speaker-specific information is important to consider, but the "accuracy" or consistency of speaker-specific information with the stimulus, should also be considered when interpreting results. In Chapter 2, the magnitude of the effect of speaker-specific information on auditory-perceptual voice ratings was greater for inaccurate versus accurate referring medical laryngeal diagnoses compared to no diagnostic information for all speaker groups (e.g., paralysis/paresis, benign lesions, control). Differences in the average perceived severity of roughness and breathiness were increased when medical laryngeal diagnoses were present versus absent, and were greatest when the presenting diagnoses

were inaccurate. This was an important finding because one large study using a national database of diagnostic codes suggested that voice clinicians are commonly exposed to both accurate and inaccurate, or non-specific referring medical laryngeal diagnoses prior to specialized voice assessment in a clinical setting.[7]

In Chapter 3, we evaluated the effect of both accurate and inaccurate case history information that suggested a particular voice disorder etiology or the absence of a voice disorder on VLS ratings. Findings indicated that clinicians were more likely to detect the presence of a visual-perceptual feature when it was suggested by the case history. However, direct comparisons of the magnitude of the effect of accurate versus inaccurate case histories on perceived severity were not possible due to the fully crossed study design. One prior study also found an effect of both fictional case histories suggesting risk or protective factors for voice disorders on vibratory VLS ratings, but the accuracy of the case histories was not considered in the study design.[6] In Chapter 3, we also found that agreement about clinical impressions among clinicians was greater when accurate versus inaccurate case histories were presented prior to VLS evaluation.

In Chapter 4, we attempted to control for the "accuracy", or consistency of the auditory cues with the VLS stimuli. Therefore, only experimental VLS exams that contained auditory cues suggesting dysphonia, and control VLS exams without laryngeal pathology that contained auditory cues suggesting no dysphonia were included in the rating task. As previously discussed, we found a significant effect of auditory cues on only one VLS parameter using a particular type of rating scale. Overall, studies examining clinical assessment tools that rely on detection of auditory- and visual-perceptual abnormalities in broader fields of medicine and voice assessment support the hypothesis that accurate information results in improved diagnostic accuracy, while

inaccurate information substantially reduces diagnostic accuracy.[8–11] A greater magnitude of the effect of inaccurate versus accurate information has also been observed in broader fields of medicine. For example, Shikino et al.[11] examined the relationship between preliminary diagnostic information and diagnostic impressions following cervical auscultation of simulated heart murmurs performed by medical students. In this study,[11] students were provided accurate, inaccurate, or no diagnostic information. Diagnostic accuracy was improved when preliminary information was present versus absent, but this difference was not statistically significant. However, when correct versus inaccurate clinical information was provided, an accurate diagnosis was generated in 87% versus 30% of trials, demonstrating the strong effect of the accuracy of information on outcomes. Therefore, the accuracy of speaker-specific information on perceptual ratings should be considered in study designs and when interpreting the results of these studies. Additionally, there are several other factors, such as the type of rating scales used to measure perceptual judgments, the severity of the stimuli, and the overall stability of the perceptual parameters being rated, that should be considered when interpreting the results of these studies.

2.3 Other Rating Task Factors

There are many different types of rating tools and scales that are used to make auditory- and visual-perceptual voice ratings in both research and clinical settings.[12–15] In Chapter 2, a 100 mm VAS labeled as “normal, no roughness/breathiness” and “severely rough/breathy” at the extreme ends of the scale were used to rate auditory-perceptual parameters of roughness and breathiness. In Chapter 4, a similar 100 mm VAS was also used to rate 9 VLS parameters and measure overall severity of laryngeal function. This type of rating scale has also been used in

other investigations that include VLS.[16,17] In Chapter 3, a 4 point ordinal scale labeled (0 = normal, 1 = mild, 2 = moderate, 3 = severe) was used to rate visual-perceptual VLS parameters. While some rating scale characteristics (e.g., ordinal versus interval, visual anchors, labels, etc.) differed between studies, all of the rating scales described above were dependent on relative evaluation. Relative evaluations require clinicians to compare their perceptions of each parameter to their internal standards for what is (ab)normal.[18] Previous studies examining the effect of speaker-specific information on perceptual voice rating tasks have also relied on rating scales that required relative evaluation.[5,6] Because relative evaluation is commonly considered to be difficult and might have affected the results of these and prior studies,[5,6] it was hypothesized that there would be a greater effect of speaker-specific information on perceptual voice ratings that required relative evaluation compared to rating scales that were not reliant on relative evaluation.[18] Therefore, in Chapter 4, we obtained VLS measures using both the 100 mm VAS and the VALI for comparison because the response scales included on the VALI are absolute (e.g., amount of excursion) are relative to another standard (e.g., percentage of non-vibrating portion of vocal fold) or relative to time (e.g., percentage of time phase that vocal fold vibration is symmetric).[14] Stroboscopic measures obtained using the VALI do not depend on a clinician's internal standard of what is considered normal.

The effects of referring medical laryngeal diagnoses and case histories suggestive of specific voice disorder etiologies in Chapters 2 and 3 were significant for several perceptual parameters using two different rating scales that required relative evaluation. However, in Chapter 4, we found a significant effect of auditory cues on ratings of only one VLS parameter when it was measured using the VALI that does not require relative evaluation. We did not find a similar effect for auditory information when VLS parameters were rated using VAS. Thus, in

contrast to our initial hypotheses, rating scale type did not appear to increase raters' susceptibility to auditory information in how they made VLS judgments. How rating scale type might be affected by other types of speaker-specific information, however, needs further investigation.

One last consideration related to rating scales pertains to how the severity of the stimuli might affect raters' judgments. For example, ratings in the middle of a unipolar rating scale (e.g., mild to moderate) have been associated with reduced rater reliability compared to those made at the extreme ends of the scale.[19,20] Ratings in this range could be more susceptible to the effect of speaker-specific information because they are less reliable among raters. The rating context and the severity of the stimuli within that experimental rating set is a another rating task factor that that might affect perceptual ratings.[3] The magnitude of the effect of speaker-specific information on auditory- and visual-perceptual voice ratings was greater for mild-moderately severe perceptual ratings versus normal or severe in two prior studies.[5,6] In those studies, the authors posited that raters' judgments were most vulnerable to biases for stimuli that were least reliable – those in the mild-moderate range.

In Chapter 3, only exams with perceptual VLS parameters that ranged from normal to moderately-severe were included. For these VLS exams, significant differences in perceived severity were observed for 2 of 3 VLS parameters. In Chapter 4, the final set of VLS exams included 4 controls and 12 experimental exams from speakers with voice complaints and structural abnormalities that affected vibratory characteristics, and who demonstrated a range of dysphonic severities (8 mild; 4 moderate-severe). Mild samples were oversampled because previous studies had shown that these types of samples might result in the strongest effects on raters' judgments. However, results from this study showed a significant interaction between

dysphonia severity and the presence of auditory cues for only one of the 9 studied VLS parameters, measured using the VALI. These differences in average VLS ratings were only significant for speakers with moderate-severe dysphonia, but not for those with mild or no dysphonia. It is unknown whether dysphonia severity was a suitable surrogate for severity of this visual-perceptual parameter in Chapter 4. In Chapter 3, severity was not evaluated because a three-way interaction was not possible due to study design. How severity might relate to the effects of other speaker-specific factors needs future study.

Although there are many potential sources of variability in auditory- and visual-perceptual VLS ratings, some auditory- and visual-perceptual parameters might be more difficult to rate than others, resulting in poorer inter-rater reliability for these parameters.[3,12,21,22] In Chapter 2, the magnitude of the effect of referring medical laryngeal diagnosis was greater for ratings of roughness versus breathiness. Consistent with the hypothesis that poorer reliability could increase raters' susceptibility to cognitive biases, we found that inter-rater reliability was also reduced for ratings of roughness compared to breathiness.

Similarly, in Chapter 3, the effect of case histories was significant for visual-perceptual VLS parameters that demonstrated lower inter-rater reliability (e.g.. posterior laryngeal appearance, mucosal wave), but not for the VLS parameter with the highest inter-rater reliability (e.g., vocal fold mobility) for visual perceptual judgments. Thus, both of the results from Chapters 2 and 3 appeared to be consistent with the hypothesis that specific parameters might be more vulnerable to cognitive biases because of their relationship with poor reliability. However, the effect of the auditory information on visual-perceptual ratings in Chapter 4 did not appear to be associated with lower levels of inter-rater reliability in rating any particular VLS parameters. In fact, inter-rater reliability was acceptable and/or even high for most visual-perceptual

parameters evaluated using both rating scales in Chapter 4. The reliability of specific parameters did not appear to be an important contributing factor when evaluating the effect of auditory information on VLS ratings in Chapter 4.

3. Future Directions

The overall findings from this series of studies suggest that some types of speaker-specific information (e.g., referring medical laryngeal diagnosis, case history suggesting a voice disorder etiology) influence auditory- and/or visual-perceptual voice rating tasks. The magnitude of the effect appeared increased when speaker-specific information was inconsistent with the stimulus presented to clinicians, and/or the overall stability of perceptual ratings for a particular parameter were reduced. However, the auditory signal that is present when VLS exams are interpreted in a clinical setting, but is commonly excluded in laboratory settings, had no and/or a minimal effect on vibratory VLS ratings when responses were measured using two different types of rating scales. As previously discussed, differences in the accuracy of the speaker-specific information provided to clinicians might also contribute to the increased magnitude of the effect observed in Chapters 2 and 3, compared to Chapter 4. However, the effect of inaccurate, or auditory information that is inconsistent with the stimulus being evaluated, on VLS ratings is unknown and requires additional investigation. Additional differences between study designs in Chapters 2, 3, and 4 might also have affected the results, and further studies to investigate these factors are warranted.

The order in which the speaker-specific information is presented to clinicians might be important to consider when comparing results from this series of studies. For example, speaker-specific information in Chapters 2 and 3 were presented before the stimulus, whereas the auditory cues were presented simultaneously with the VLS exam in Chapter 4. It is possible that

the effect of auditory information about dysphonia severity might differ if auditory cues are presented prior to, rather than simultaneously with the VLS exam. The order that information is presented can influence its perceived relevance.[9] Previous research in broader fields of medicine also suggests that information presented first is used to generate preliminary hypotheses and expectations that can affect how auditory and visual information is evaluated to make medical diagnoses.[8,9,11] Differences in the order that speaker-specific information was presented might have affected the results from Chapter 4 and should be considered in future study designs. This is also an important consideration because clinicians are exposed to a speaker's voice quality not only during the VLS exams, but also prior to evaluating VLS exams in a clinical setting.

Two additional factors that should be considered are the amount of cues that are provided by speaker-specific information and/or their relevance for making perceptual ratings or other types of clinical judgments. For example, case histories in Chapter 3 were intended to strongly suggest a specific voice disorder etiology, but they also provided clinicians with information about a speaker's symptom severity, medical history, and occupation, among other factors (e.g., symptom onset, course, and type, dietary factors, hydration). Similarly, while the audio signal presented to clinicians in Chapter 4 likely provided clinicians with cues about dysphonia severity, additional auditory cues about speech or voice characteristics (e.g., voice quality, pitch, loudness, tremor, voice breaks, etc.) might also affect the results. These factors also should be considered in future studies investigating these potential sources of cognitive biases on perceptual judgments.

The speaker-specific information provided to clinicians in Chapters 2, 3 and 4 (e.g., referring medical laryngeal diagnosis, case history, etc.) are generally considered to be irrelevant for making auditory- and/or visual-perceptual voice assessment. Therefore, the systematic

differences in perceptual ratings commonly attributed to cognitive bias, defined as systematic deviations from a normative response, might be considered "errors".[4] Indeed, the effects of speaker-specific information on average perceptual ratings observed in Chapters 2 and 3, and in prior studies have been attributed to cognitive biases.[5,6] However, there were also positive effects of speaker-specific information suggesting a particular voice disorder etiology on the detection of visual-perceptual abnormalities in Chapter 3. Whether differences in detection of perceptual features and/or differences in their perceived severity might be clinically meaningful also requires further investigation. For example, does a 20 mm difference in perceived severity of roughness indicate adequate improvement with voice treatment(s)? Or, does a 20 mm increase in perceived severity of non-vibrating portion of a vocal fold lead to surgical management versus recommendations for behavioral intervention or a specific medical laryngeal diagnosis? Although some of these differences in detection or the perceived severity of abnormalities could theoretically lead to differences in diagnosis and/or treatment recommendations, this also requires further investigation.

One additional caveat is that speaker-specific information is considered irrelevant for making perceptual ratings, but is often important for making clinical decisions about diagnosis and/or treatment. In Chapter 3, a case history that suggested the absence of voice symptoms and normal laryngeal function uniquely contributed to clinicians' decisions to recommend no treatment versus treatment, above and beyond the VLS exam evaluated with this case history. This finding suggested a unique effect of speaker-specific information on clinical decisions about treatment recommendations. Therefore, different types of speaker-specific information might contribute to clinical decisions about diagnosis and/or treatment directly and/or indirectly (refer to figure 1 in Chapter 4). Future study designs should consider these complex relationships, and

the relevance of information obtained from different clinical voice assessment tools for making decisions about clinical diagnosis and treatment.

Lastly, the relative importance of findings obtained from various clinical assessment tools likely varies by voice disorder etiology. Studies in Chapters 2, 3, and 4 included primarily speakers with diagnoses of benign lesions and/or vocal fold paralysis/paresis, and excluded speakers with malignancies and/or airway compromise. Future studies investigating the effect of speaker-specific information on auditory- and visual-perceptual ratings and/or clinical decisions about diagnosis and treatment should include speakers with additional voice disorder etiologies. Inclusion of voice disorders that might represent challenging and important differential diagnoses (e.g., spasmodic dysphonia versus muscle tension dysphonia, presbylaryngis versus hypophonia secondary to dysarthria, benign versus malignant lesions) might be especially important to investigate. Understanding the sources of variability in perceptual voice assessment measures and other types of clinical judgments is important for supporting the validity of perceptual measures, including VLS, for determining the presence, severity, or nature of a voice disorder in clinical and laboratory settings. It is hoped that these investigations will help us better serve these patients using evidence-based approaches to both assessment and treatment.

References

1. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am J Speech Lang Pathol*. 2018;27(3):887-905.
2. Behrman, Alison. "Common Practices of Voice Therapists in the Evaluation of Patients." Journal of Voice, vol. 19, no. 3, 2005, pp. 454–469.
3. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *J Speech Hear Res*. 1993;36(1):21.
4. Kahneman D. Kahneman, D., Slovic, P., & Tversky A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.
5. Eadie T, Sroka A, Wright DR, Merati A. Does Knowledge of Medical Diagnosis Bias Auditory-Perceptual Judgments of Dysphonia?(Report). 2011;25(4):420.
6. Teitler N. Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. *J Voice*. 1995;9(1):95-105.
7. Cohen SM, Kim J, Roy N, Wilk A, Thomas S, Courey M. Change in diagnosis and treatment following specialty voice evaluation: A national database analysis. *Laryngoscope*. 2015;125(7):1660-1666.
8. Hatala R, Norman GR, Brooks LR. Impact of a clinical scenario on accuracy of electrocardiogram interpretation. *J Gen Intern Med*. 1999;14(2):126-129.
9. Leblanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. *Acad Med*. 2002;77(10 Suppl):S67-69.
10. Sauder C, Nevdahl M, Kapsner-Smith M, Merati A, Eadie T. Does the accuracy of case history affect interpretation of videolaryngostroboscopic exams? *The Laryngoscope*. 2020;130(3):718-725.
11. Shikino K, Ikusaka M, Ohira Y, et al. Influence of predicting the diagnosis from history on the accuracy of physical examination. *Adv Med Educ Pract*. 2015;6:143-148.
12. Bonilha HS, Desjardins M, Garand KL, Martin-Harris B. Parameters and Scales Used to Assess and Report Findings From Stroboscopy: A Systematic Review. *Journal of Voice*. 2018;32(6):734-755.
13. Nawka T, Konerding U. The interrater reliability of stroboscopy evaluations. *J Voice*. 2012;26(6):812.e1-10.
14. Poburka BJ, Patel RR, Bless DM. Voice-Vibratory Assessment With Laryngeal Imaging (VALI) Form: Reliability of Rating Stroboscopy and High-speed Videoendoscopy. *J Voice*. 2017;31(4):513.e1-513.e14.
15. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J*. 1987;66(7):289-296.
16. Roy N, Barton ME, Smith ME, Dromey C, Merrill RM, Sauder C. An in vivo model of external superior laryngeal nerve paralysis. *Laryngoscope*. 2009;119(5):1017–1032.
17. Heller A, Tanner K, Roy N, et al. Voice, Speech, and Laryngeal Features of Primary Sjögren's Syndrome. *Annals of Otology, Rhinology & Laryngology*. 2014;123(11):778-785.
18. DeCastellarnau A. A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity*. 2018;52(4):1523-1559.

19. Law T, Kim JH, Lee KY, et al. Comparison of Rater's reliability on perceptual evaluation of different types of voice sample. *J Voice*. 2012;26(5):666.e13-21.
20. Eadie TL, Kapsner-Smith M. The Effect of Listener Experience and Anchors on Judgments of Dysphonia. *J Speech Lang Hear Res*. 2011;54(2):430-447.
21. Chan KMK, Yiu EM-L. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45(1):111-126.
22. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20(4):527-544.