Original Paper

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2018;70:156–164
DOI: 10.1159/000492219

Published online: August 29, 2018

The Effect of the Speech Task Characteristics on Perceptual Judgment of Mild to Moderate Dysphonia: A Methodological Study

Jérôme R. Lechien[a–c] Dominique Morsomme[d] Camille Finck[b, d] Kathy Huet[b] Véronique Delvaux[b] Myriam Piccaluga[b] Bernard Harmegnies[b] Sven Saussez[a, c]

[a]Laboratory of Anatomy and Cell Biology, Faculty of Medicine, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium; [b]Laboratory of Phonetics, Faculty of Psychology, Research Institute for Language Sciences and Technology, University of Mons (UMons), Mons, Belgium; [c]Department of Otorhinolaryngology and Head and Neck Surgery, RHMS Baudour, EpiCURA Hospital, Baudour, Belgium; [d]Department of Speech Therapy, Voice Unit, University of Liège, Liège, Belgium

Keywords
Voice · Perceptual assessment · Speech task

Abstract
Objective: To study the differences in perceptual ratings of mild and moderate dysphonia related to the speech task, and their impact on intrarater and interrater reliabilities. ***Patients and Methods:*** Voice recordings of 15 outpatients with mild or moderate dysphonia related to laryngopharyngeal reflux were presented to 6 female experienced judges blinded to the clinical state of the patients. From these, the GRBASI (Grade, Roughness, Breathiness, Asthenia, Strain, and Instability) evaluations were performed on connected speech and sustained vowel of the pretreatment voice recordings and absolute agreement, and both intrarater and interrater reliabilities were assessed. ***Results:*** The average GRBASI scores were significantly worse when performed on sustained vowel. Intrarater reliability substantially varied according to the judge and the task. Good interrater reliability was broadly found for the evaluations of all GRBASI components irrespective of the speech task. Concerning agreement, we only found absolute agreement between judges for G and R items assessed on text. ***Conclusion:*** Average grade of perceptual voice impairment, intrarater reliability, and agreement vary according to the speech task.

Introduction

Voice quality studies usually involve subjective and objective assessments to get an overall idea of the patient's vocal function. Most of the time, subjective evaluations include self-report questionnaire and perceptual judgment of the patient voice. The perceptual assessment is usually performed with standardized tools providing reliable scores of some voice characteristics. The most common perceptual tool used around the world still remains the Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) scale that was initially introduced by Hi-

B.H. and S.S. contributed equally to this work and should be regarded as *joint last authors*.

E-Mail karger@karger.com
www.karger.com/fpl

Jérôme R. Lechien, MD, PhD, MS
Laboratory of Anatomy and Cell Biology, Faculty of Medicine
University of Mons (UMONS), Avenue du Champ de mars, 6
BE–7000 Mons (Belgium)
E-Mail jerome.lechien@umons.ac.be

rano et al. [1] in 1981 and completed (with Instability, GRBASI) by Dejonckere et al. [2] in 1996. In the current literature, an important methodological variability exists concerning the completion of the GRBASI grading, since it may be based on sustained vowel [3, 4], connected speech [5], conversation, counting task [6], or reading text [7, 8]. These voice samples can be used in another perceptual tool such as the Consensus Auditory-Perceptual Evaluation of Voice, which includes different tasks, i.e., vowels, connected speech, and conversation [9]. The sustained vowel has the advantage of being easier to elicit and still being unaffected by the articulation of the subject [10, 11] but this production model is unnatural and not representative of the daily speaking voice [12]. In contrast, continuous speech is more natural but some speaking characteristics (articulation, speech rate) may substantially influence the assessment of the voice quality [13]. With regard to the debate about both the reliability and validity of one or the other vocal sample used, a few authors use composite methods alternating continuous speech to assess certain GRBASI components (e.g., roughness, breathiness, and instability) and sustained vowel for other components (i.e., asthenia and strain) [14–16]. Moreover, in the last decades, a few researchers have studied the impact of the choice of the task on the results of the perceptual voice quality evaluations and they found mixed results, especially concerning the interrater reliability and agreement between judges [13, 17–19]. Thus, some data supported that sustained vowels are rated significantly more dysphonic than continuous speech [20], while other did not find a significant difference [18]. The controversy also concerns the impact of speech task on the interrater reliability since some authors did not find significant differences in interrater reliability between dysphonia severity ratings of sustained vowels and continuous speech [13, 20, 21], while others supported that the counting task and/or sustained /a/ phonation are the optimal tasks for perceptual voice judgment with regard to interrater reliability [17]. Most of the studies selected perceptual moderate or severe dysphonia samples to demonstrate a task's effect, but they did not compare the absolute agreement, intra- and interrater reliabilities of mild or moderate dysphonia that often remain harder to grade than severe dysphonic voice [22, 23].

The first objective of this study is to examine the differences of perceptual ratings of mild and moderate dysphonia related to the speech task. The second objective is to highlight a potential effect of the speech task on the agreement, and intra- and interjudge reliabilities.

Materials and Methods

Subject Characteristics

Eighty outpatients with a clinical diagnosis of laryngopharyngeal reflux disease (LPRD) were recruited from September 2014 to May 2016 in both the EpiCURA Hospital and Liège University Hospital. The study has been approved by the institutional review board of CHU de Liège (ref. B707201524621) and informed consent was obtained from all patients. The LPRD diagnosis was made using the reflux symptom index and the reflux finding score respecting the thresholds of Belafsky (reflux symptom index >13 and reflux finding score >7) [24]. From these patients, based on blinded speech recording regarding the moment of the recording (pre- vs. post-treatment), an experienced speech therapist selected patients presenting a mild or moderate perceptual dysphonia. The determination of mild and moderate dysphonia was based on the grade of dysphonia (G1 = mild; G2 = moderate) of the GRBASI scale. So, 15 patients were included in this study (6 female and 9 male) with an average age of 56.47 years (age range: 19–72 years). The mean voice handicap index score at baseline was 27.07 ± 19.06. The characteristics of the cohort are available in Table 1. Subjects with vocal fold lesions such as polyp, nodule and edema, or cofactors influencing voice quality (i.e., infections in the last month, tobacco, alcohol, etc.) were excluded from the study. All patients received a diet and behavioral advice and medical treatment for 3 months (pantoprazole, 20 mg twice daily).

Patient Voice Recordings

At baseline and after 3 months of treatment, patients were invited to read a phonetically balanced text and instructed to produce the vowel /a/ at a comfortable pitch and loudness level, three times for a time corresponding to the maximal phonation time. The voice recordings were carried out by the same practitioner during the consultation in an anechoic chamber with a high-quality microphone (Sony PCM-D50; New York, NY, USA) placed at a distance of 30 cm from the patient's mouth. Recordings were set in a sound matrix in a blinded manner according to the moment (before or after therapy) of the recording. In other words, the judge did not know the status of the patient associated with the recording.

Judges and Perceptual Assessment

The perceptual voice evaluation (GRBASI) of hoarse patients was performed by a jury of experienced judges (5 experienced speech therapists and 1 physician, all female, with ≥4 years of full-time practice) who completed a test-retest procedure from the recording matrix. Each judge was invited to characterize the patient's voice following the classical definitions of each component of GRBASI (from 0 = no alteration to 3 = severe alteration). The judges were blinded with regard to the therapeutic time of the voice recording (voice sample from baseline vs. after treatment). Once the task was done, we collected and concentrated the rest of the analysis on the perceptual evaluations of the pretreatment time. Thus, the characterization of the voice sample as mild or moderate dysphonia was only based on the recordings performed at baseline. All of these judges were specialized in voice disorders and had a regular practice. The experience of the judges, the years of experience, and the practice place (hospital vs. at home) are given in Table 2. They were familiar with the utilization of the GRBASI scale.

Table 1. Characteristics of the cohort

Speaker	Gender	Age, years	Ethnicity	RSI pre	RSI post	RFS pre	RFS post	LPRD state[a]	Dysphonia	VHI
1	M	50	Caucasian	18	1	13	13	resistant	mild	8
2	M	56	Caucasian	31	0	11	0	responder	mild	35
3	F	59	Caucasian	20	24	9	3	resistant	mild	8
4	F	72	Caucasian	17	3	12	2	responder	mild	20
5	F	45	Caucasian	35	7	12	6	responder	mild	29
6	F	41	Caucasian	39	11	14	2	responder	mild	25
7	M	55	Caucasian	19	5	11	2	responder	mild	29
8	M	19	Caucasian	17	9	9	3	responder	mild	2
9	M	71	Caucasian	17	2	22	16	resistant	moderate	57
10	M	64	Caucasian	33	13	15	1	responder	moderate	72
11	M	71	Caucasian	21	13	12	10	resistant	mild	28
12	F	43	Caucasian	16	9	14	4	responder	mild	5
13	F	84	Caucasian	16	11	11	7	resistant	mild	21
14	M	50	Caucasian	28	22	11	8	resistant	mild	27
15	M	67	Caucasian	18	18	8	5	resistant	mild	40

F, female; LPRD, laryngopharyngeal reflux disease; M, male; RFS, reflux finding score; RSI, reflux symptom index; VHI, voice handicap index. [a] The patient was considered as responder with an RSI <14 and an RFS <8 after 3 or 6 months of treatment.

Table 2. Characteristics of the judges

Judges				Activity characteristics		
No.	degree	age, years	YT	places	speech ass.	speech trt.
1	CCC-SLP	30	9	hospital, at home	yes	yes
2	MD	36	12	hospital	yes	follow-up
3	CCC-SLP	35	13	at home	no	yes
4	CCC-SLP	33	12	at home	no	yes
5	CCC-SLP	26	4	hospital	yes	yes
6	CCC-SLP	29	9	hospital, at home	yes	yes

ass, assessment; trt, treatment; YT, years of training.

Empirical Approach

Severity Rating according to the Task

To study the differences of perceptual ratings related to the speech task, we compared, for each judge, the mean values of each component of GRBASI according to the task.

Intrarater Reliability

To perform the intrarater reliability analysis, the judges were invited to assess the GRBASI scores on the phonetically balanced text and the sustained vowel (both on the entire sample) at an initial time (d0) and 7 days after the first evaluation (d7) (test-retest procedure). If possible, they were encouraged to carry out the voice assessment according to the sample (text vs. sustained vowel) with a minimum of a 24-h gap between the recording types to limit the influence of one analysis on the other. Thus, we can consider that there were 4 sessions: (i) evaluation of sustained vowel (d0), (ii) evaluation of connected speech (d1), (iii) re-evaluation of sustained vowel (d7), and (iv) re-evaluation of connected speech (d8). This test-retest reliability was assessed for each item of the GRBASI scale, using a correlation analysis for each item. We also conducted a correlation study between the values of correlation coefficients and the number of experience years of each judge.

Interrater Reliability and Agreement

An interjudge reliability analysis was made for components of the GRBASI scale in both speech tasks with Kendall's W and agreement was assessed by the Friedman test. Precisely, Kendall's W was used to assess the similarity between judges in the classification of GRBASI items (and the related interrater reliability), while the Friedman test was used to evaluate the similarity (absolute agreement) of the values of the scores given by the judges.

Table 3. Mean values of the GRBASI components according to the voice sample

	Mean (text)	Mean (/a/)	*z*	*p*
Grade	0.95±0.61	1.51±0.65	–2.36	0.019
Roughness	0.86±0.51	1.33±0.73	–1.50	0.137
Breathiness	0.31±0.38	0.69±0.55	–2.20	0.029
Asthenia	0.33±0.27	0.66±0.45	–2.09	0.041
Strain	0.72±0.52	0.99±0.42	–1.88	0.061
Instability	0.69±0.49	1.27±0.64	–2.43	0.015

According to the Mann-Whitney test, the mean values of grade of dysphonia, breathiness, asthenia, and instability were significantly higher when the perceptual analyses were performed on sustained /a/.

Statistical Analysis

Statistical analysis was performed using the Statistical Package for the Social Sciences for Windows (SPSS version 22.0; IBM Corp. Armonk, NY, USA). The comparison of the mean values of all components of GRBASI at baseline according to the speech sample for each judge was made with the Mann-Whitney test. The assessment of the reliability based on the test-retest correlation was made using Spearman's rank correlation coefficient. The interrater reliability and the agreement analyses were made for each item with Kendall's W and the Friedman test, respectively. A level of significance of 0.05 was adopted.

Results

Severity Assessment according to the Task

The mean values of the GRBASI components according to the voice sample are available in Table 3. According to the Mann-Whitney test, the ratings of grade of dysphonia ($p = 0.019$), breathiness ($p = 0.029$), asthenia ($p = 0.041$), and instability ($p = 0.015$) were significantly higher when the voice quality was assessed on sustained vowel than on the text.

Intrarater Reliability

The descriptive statistics of intrarater reliability for the voice assessments carried out on text and sustained vowel are available in Table 4. Concerning the evaluations performed on the text, from the 6 judges, 2 have similarly assessed all GRBASI items during the two sessions (judges 5 and 6) and reported good intrarater reliability (correlation coefficient from 0.999 to 0.614; $p < 0.026$). Judge 4 reported poor intrarater reliability of all GRBASI components (correlation coefficients from 0.098 to 0.403; $p > 0.050$). The intrarater reliability of other judges depended of the items of GRBASI (judges 1, 2, and 3; Table 4).

Concerning the evaluations performed on sustained vowel, we also found 2 judges with high intrarater reliability regarding all GRBASI items (judges 1 and 6; correlation coefficients from 0.590 to 0.928; $p < 0.021$). Judges 3 and 4 had poor intrarater reliability (correlation coefficients from 0.052 to 0.441; $p > 0.05$). The intrarater reliability of other judges depended of the items of GRBASI (judges 2 and 5; Table 4).

Overall, when the correlation coefficient was significant for the test-retest procedure of one task, we statistically observed a similar trend for the other task. Thus, correlation coefficients for the GRBASI evaluations on text were 0.495, 0.649, 0.514, 0.643, 0.540, and 0.572, respectively ($p < 0.05$). The same correlation coefficients for the GRBASI evaluations on sustained vowel were 0.566, 0.482, 0.348, 0.632, 0.527, and 0.531 ($p < 0.05$).

Moreover, our correlation study between the values of the correlation coefficients for each item of the GRABSI and the number of experience years did not reveal significant relationship.

Interrater Reliability

According to Kendall's W, we found good interrater reliability between judges for all items of the GRBASI scale irrespective of the speech sample ($p < 0.04$; Table 5). Concerning agreement, according to the Friedman analysis, we only found good agreement for G ($p = 0.841$) and R ($p = 0.423$) items assessed on text. The mean values of the GRBASI components (± standard deviation) are described in Table 6.

Discussion

The perceptual evaluation of the voice quality is a complex cognitive process implying interrelated known and unknown mechanisms such as training and experience of the judge, the knowledge of the patient's medical history, the used tool to assess the voice quality, and the type of speech stimuli [17, 25, 26]. Most of the studies interested in these mechanisms used heterogeneous or moderate to severe dysphonic speech samples, inferring their results to all states of dysphonia. In this study, we assessed the impact of the speech task on the judgment of samples of mild to moderate dysphonic voice.

Table 4. Intrarater reliability for the voice assessments carried out on text and sustained vowel

Judges	G (Grade): test-retest		R (Roughness): test-retest		B (Breathiness): test-retest		A (Asthenia): test-retest		S (Strain): test-retest		I (Instability): test-retest	
	Spearman	*p* value	Spearman	*p* value	Spearman	*p* value	Spearman	*p* value	Spearman	*p* value	Spearman	*p* value
Sample: text												
1	0.401	0.175	0.608	0.027	0.796	0.001	0.882	0.001	0.541	0.056	0.589	0.034
2	0.909	0.001	0.779	0.002	0.178	0.561	0.999	0.001	0.88	0.001	0.543	0.055
3	0.133	0.664	0.621	0.024	0.426	0.147	0.123	0.689	0.212	0.489	0.673	0.012
4	0.143	0.641	0.403	0.172	0.298	0.323	0.098	0.751	0.161	0.6	0.389	0.189
5	0.627	0.022	0.785	0.001	0.778	0.002	0.999	0.001	0.871	0.001	0.636	0.019
6	0.760	0.003	0.698	0.008	0.626	0.022	0.755	0.003	0.652	0.016	0.614	0.026
Sample: sustained vowel												
1	0.928	0.001	0.876	0.001	0.769	0.001	0.894	0.001	0.590	0.021	0.788	0.001
2	0.499	0.058	0.383	0.159	0.163	0.562	0.990	0.001	0.734	0.001	0.670	0.009
3	0.168	0.584	0.052	0.865	0.150	0.624	0.099	0.749	0.102	0.741	0.061	0.844
4	0.255	0.360	0.226	0.418	0.190	0.499	0.275	0.322	0.441	0.100	0.127	0.651
5	0.615	0.015	0.476	0.073	0.045	0.873	0.627	0.012	0.705	0.003	0.751	0.001
6	0.928	0.001	0.876	0.001	0.769	0.001	0.894	0.001	0.590	0.021	0.788	0.001

The intrarater reliability analyses were performed with the Spearman correlation test.

Table 5. Interrater reliability (text and sustained vowel)

		G	R	B	A	S	I
Text	Kendall's W	0.582	0.514	0.575	0.301	0.470	0.543
	p value	<0.001	<0.001	<0.001	0.041	0.001	<0.001
	Friedman	0.032	0.076	0.184	0.344	0.235	0.191
	p value	0.841	0.423	0.035	<0.001	0.009	0.029
Vowel	Kendall's W	0.734	0.771	0.515	0.418	0.351	0.603
	p value	<0.001	<0.001	<0.001	<0.001	<0.001	0.003
	Friedman	0.326	0.498	0.314	0.597	0.399	0.235
	p value	<0.001	<0.001	<0.001	<0.001	<0.001	0.003

The interrater reliability analyses were performed with both Friedman and Kendall's tests. According to our analysis, only the evaluations of grade of dysphonia and roughness reported high interrater reliabilities when these parameters were evaluated on text. GRBASI, Grade, Roughness, Breathiness, Asthenia, Strain, and Instability.

Table 6. The GRBASI components (mean ± SD) according to the judge and the voice sample

Judges	G (Grade)		R (Roughness)		B (Breathiness)		A (Asthenia)		S (Strain)		I (Instability)	
	text	vowel	text	vowel	text	vowel	text	vowel	text	vowel	text	vowel
1	0.92±0.64	1.53±0.52	1.00±0.58	1.87±0.83	0.62±0.77	1.27±1.03	0.46±0.66	0.73±0.70	0.38±0.77	1.13±0.74	0.69±0.63	1.40±0.91
2	0.77±0.92	1.13±0.83	0.62±0.96	0.87±0.74	0.15±0.38	0.27±0.59	0.00±0.00	0.00±0.00	0.38±0.65	0.13±0.52	0.31±0.48	0.80±0.68
3	1.08±0.86	1.47±0.64	0.85±0.69	0.87±0.92	0.31±0.48	0.67±0.62	0.08±0.28	0.47±0.74	1.00±0.82	0.80±0.56	0.62±0.87	1.13±0.60
4	0.92±0.86	1.47±0.92	0.85±0.80	1.40±0.91	0.23±0.44	0.53±0.74	0.54±0.66	1.40±0.74	0.85±0.56	1.40±0.74	1.00±0.71	1.20±0.78
5	1.00±0.82	1.40±0.74	0.92±0.76	1.33±0.82	0.15±0.38	0.40±0.63	0.15±0.38	0.07±0.26	1.00±0.71	1.33±0.82	0.85±0.69	1.33±0.90
6	1.00±0.58	2.07±0.96	0.92±0.49	1.67±0.82	0.38±0.51	1.00±0.85	0.77±0.60	1.27±1.10	0.69±0.75	1.13±0.74	0.69±0.63	1.73±1.03

Severity of the Perceptual Evaluations according to the Speech Sample

Firstly, our results show significantly higher values of GRBASI components assessed on sustained vowel than those assessed on connected speech. A similar effect of speaking task was found in the study of Wolfe et al. [27] who reported significantly higher ratings of the grade of dysphonia assessed on sustained vowel than those assessed on connected speech in both dysphonic and healthy subjects. The study of Lu et al. [17] showed identical findings of the ratings of roughness and breathiness, while Zraick et al. [18] also reported more severe judgments of dysphonia graded on sustained vowel. These observations contrast with methodological studies of de Krom [10] and Revis et al. [11] who did not report any discrepancy in the evaluations of the perceptual voice quality according to the task. These divergent results observed among these studies could be the consequence of discrepancies in the composition of their voice samples (especially the severity of the dysphonia) since it has been demonstrated that the assessment of patients with higher dysphonia scores may reduce the variability of the score severity related to the task [12, 13]. Among the trials using sustained vowels, some authors removed the onset/attack of the signal while others assessed the perceptual voice quality on the entire signal. Regarding the fact that the voice onset may contain important data for the perceptual assessment [12, 28], removing of the voice onset could reduce the severity of the perceptual grade attributed to the voice sample. In this, paper, although we removed the onset of the sustained vowel, we show higher scores of GRBASI components when based on sustained vowel than when based on connected speech. Thus, the impaired voice quality is still perceptible along the "steady portion" of the vowel. In addition, even if connected speech is more representative of the day-to-day speech, some factors related to the intrinsic speech features of the patient (i.e., speed of the reading, language accent, pitch, loudness, articulation) may potentially mask embedded voice alterations in the voice quality assessment procedure [17]. As reported in a previous study [17], our results support that speech tasks influence the perceived voice quality of the subject and, de facto, the ability of the judge to reliably judge the voice.

Intrarater Reliability according to the Task

We found that the values of correlation coefficients of the test-retest reliability significantly depend on the judge and they are broadly stable over the speech tasks, for every judge. In other words, judges with poor values of correlation coefficients on sustained vowel had also poor values for the evaluations made on text and vice versa. Similar results had already been observed in the current literature [13, 29, 30]. Recent publications argued that the most important findings accounting for intrarater reliability remain experience and professional background of the judges [18, 31]. Two judges (judges 3 and 4) presented poor intrarater reliability, which cannot be explained by the years of practice (12 and 13 years). However, these judges have a different practice in comparison with the other judges in whom the values of correlation coefficients are better. They have a private practice (at home) mainly focused on therapeutic management (based on instructions of an initial report) and they do not carry out the initial speech assessment. Moreover, they do not work in a hospital where, in Belgium, there is a higher heterogeneity of the dysphonic cases. These findings represent a kind of lack of experience regarding varied clinical practice. We may also postulate that other factors may explain these results and include the mental or psychic condition of the judge, potential fatigue or attention lapse at the moment of the evaluations, or the voice types listened in their daily consultation before the beginning of the assessment [13, 32].

In our intrarater reliability analysis, we observed a few different values of correlation coefficients varying in accordance with the task. On this point, controversial results exist in the current literature. Indeed, some authors found higher reliability of voice quality assessments based on sustained vowel than connected speech, especially breathiness and roughness ratings [10], while others supported contrary findings [13]. In our study (Table 6), as in the study of Law et al. [13] composed of a majority of mild and moderate dysphonia, we find a few higher values of correlation coefficients in the test-retest procedure of text compared to sustained vowel task. Three hypotheses may explain these findings. Firstly, the common practice taught in our speech therapist schools mainly promotes an assessment method of voice quality based on continuous speech that leads to better trainings and habits with connected speech samples explaining a more comfortable assessment of voice quality with continuous speech task. Secondly, the rating of the individual components of the perceptual voice quality (G, R, B, A, S, and I) may be tougher on a "static" vocal task (sustained vowel) unlike the revealing nature of connected speech. It has been shown that connected speech provides more apparent and dynamic signals than sustained vowel (related to the greater laryngeal muscular movement and the impact on the pitch, loudness, and tessitura), which remains more reliable to judge [10, 13, 33]. Interestingly, the recent paper of Gerratt et al. [33] pro-

posed that the perceptual differences between sustained vowel and continuous speech derived from the variability of voice source across segmental and prosodic contexts and not from variations of the vocal cord vibration in the quasi-steady portion of the vowel [21]. Thirdly, connected speech is more frequently heard in daily life of the judge, conferring a greater auditory experience and memory of the dysphonic voice. Concerning our values of correlation coefficients, they are lower than in the studies of Freitas et al. [25] and Law et al. [13]. A possible explanation is the discrepancies between our studies related to the severity of the overall dysphonia since our group is mainly composed of mild to moderate dysphonia.

Interrater Reliability and Agreement according to the Task

We observed good interrater reliability between judges for all items of the GRBASI scale irrespective of the speech sample. We also found good agreement only for G and R evaluations assessed on reading text. These observations corroborate some previous studies supporting better interjudge reliability and agreement on G, R and B items [34]. In contrast, our interrater reliability values (Spearman correlation coefficients) remain slightly lower than those described in studies based on moderate-to-severe dysphonia [13]. Firstly, it has been shown that severe dysphonia is rated more reliably between judges (intra- and interrater reliabilities) than mild and moderate dysphonia [13]. Thus, we think that the lower interrater reliability values found in our study are primarily due to the lower perceptual voice quality impairments characterizing our samples. Some papers supported our explanation since the interrater reliability for the levels of the severity setting in the mid-range of the GRBASI scale often had the greatest variation [12, 13].

Secondly, our lower values of correlation coefficients may be related to some intrinsic and extrinsic factors influencing the voice quality assessment. Intrinsic factors may include the mental representation of normal and dysphonic voice (that may vary from one judge to another), a judge's memory, attention to the task, education of the judge, and the gender of the judge [19, 20, 35, 36]. To avoid the impact of this last factor, we have chosen female judges to reduce the gender effect in the assessment of the voice quality. Extrinsic factors consist of acoustic context, type of listening tasks, speaking rate, pitch glide of the production, and the length of speech samples [17, 26]. Thus, it seems that longer segments yield lower rater reliability values related to fatigue, inattentiveness and distraction [17]. To avoid this bias, as recommended [17], our sustained vowel samples consisted of a segment duration of no more than 5 s by deleting the beginning of the signal while our text consisted of one sentence from a phonetically balanced text. It is also important to mention that removing the beginning of the signal could increase the difficulty of the task evaluation since some signal "imperfections" are identified at the beginning of speech. The assessment of voice quality also requires good perceptual sensitivity to the items being rated [37]. To reduce this bias, we preliminarily used a standardized tool with clear definition of the components assessed by this tool. It had been ensured that all of our judges were familiar with the GRBASI tool since they usually used this scale in their daily practice. The clinical experience is a contested extrinsic factor since some studies suggested that it is even more the professional background that impacts the perceptual voice evaluations and the interrater reliability than the number of years of clinical experience [35]. On the contrary, other authors suggested that more experienced judges are more reliable raters [1, 18, 31, 37]. We observed that the criteria used to define the "experience" of the judges vary between publications, explaining the mixed results. Our results fail to find a significant effect of the years of experience but, as mentioned above, support a potential effect of the work site/context and, indirectly, the variability of both the practice and the patient profiles.

We consider the blinded assessment of the voice quality of our patients as a strength of our study. Indeed, the judges did not know both the state of the listened voice and the clinical findings at the moment of the evaluation that limited the potential bias related to the knowledge of the clinical state underlying the sample [18]. Another strength of this study concerns the single disease (LPRD) on which the analyses were done. Thus, diseases with a high range of severity of dysphonia, or the presence of various diseases in a heterogenic group, may bias (and reduce) the intra- or interjudge reliabilities of the perceptual voice quality evaluations. All patients included in this study had mild to moderate LPRD, allowing a high consistency of the assessed recordings.

Some limitations characterize our study, which may lead to directions for future research. Firstly, the low number of patients (and analyzed voice samples) limits the statistical power of our results and this could therefore lead to an underestimation of some tendencies. But, on the other hand, our test-retest procedure required concentration of the judges and the increase of the speech samples could decrease the efficiency of the task realization. Secondly, it remains perilous to generalize the results of our judges to other groups of judges regarding

their characteristics related to our research and clinical environment (i.e., features of the language of the country, school formation, and professional background). Moreover, we did not control many independent variables including speaking rate, pitch variation, pauses between syllables or words, etc., which may impact the final result. However, it may be interesting to study the reliability of voice quality evaluations of other groups of judges including general otolaryngologists and phoniatricians, who are often the first professionals confronted with making a voice perceptual judgment. Thirdly, we did not consider the lack of a control group as an important weakness since we less frequently assess normal voice in our practice. Moreover, the aim of this study focused on the specific behavior of judges assessing mild to moderate voice disorder according to the speech task.

Conclusion

Our study contributes to the literature suggesting a significant impact of the speech task characteristics on the perceptual voice quality raters. As other results previously, but especially in mild to moderate dysphonia, our results highlight the need to standardize the methodological approach to assess the perceptual voice quality, because otherwise comparison between studies still remains difficult regarding the methodological discrepancies between studies. Future studies are needed to determine the most suitable speech stimuli for perceptual ratings to yield to better intra- and interjudge reliability in all dysphonia degrees of severity.

Acknowledgment

We thank the American Journal Expert for the proofreading of the paper, ARC Grant of the French Speaking Community of Belgium, and speech therapists and physician constituting the jury (Ondine Genat, CCC-SLP; Aude Lagier, MD; Charlotte Simon, CCC-SLP; Virginie Roig-Sanchis, CCC-SLP; Lorraine Lieffrig, CCC-SLP; Eva Ficarrotta, CCC-SLP).

Disclosure Statement

The authors declare that they have no conflict of interest.

References

1 Hirano M, Hibi S, Teresawa R, Fujiu M. Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates of dysphonia. J Phonetics. 1986;14:445–56.
2 Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier-Buchman L, Millet B. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. Rev Laryngol Otol Rhinol (Bord). 1996;117(3):219–24.
3 Dedivitis RA, Barros AP, Queija DS, Alexandre JC, Rezende WT, Corazza VR, et al. Interobserver perceptual analysis of smokers voice. Clin Otolaryngol Allied Sci. 2004 Apr; 29(2):124–7.
4 Santos KW, Scheeren B, Maciel AC, Cassol M. Vocal Variability Post Swallowing in Individuals with and without Oropharyngeal Dysphagia. Int Arch Otorhinolaryngol. 2015 Jan; 19(1):61–6.
5 Sardesai MG, Merati AL, Hu A, Birkent H. Impact of patient-related factors on the outcomes of office-based injection laryngoplasty. Laryngoscope. 2016 Aug;126(8):1806–9.
6 Silva LF, Gama AC, Cardoso FE, Reis CA, Bassi IB. Idiopathic Parkinson's disease: vocal and quality of life analysis. Arq Neuropsiquiatr. 2012 Sep;70(9):674–9.
7 Nemr K, Simões-Zenari M, Cordeiro GF, Tsuji D, Ogawa AI, Ubrig MT, et al. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. J Voice. 2012 Nov;26(6):812.e17–22.
8 Lechien JR, Delvaux V, Huet K, Khalife M, Fourneau AF, Piccaluga M, et al. Phonetic Approaches of Laryngopharyngeal Reflux Disease: A Prospective Study. J Voice. 2017 Jan;31(1):119.e11–20.
9 Mozzanica F, Ginocchio D, Borghi E, Bachmann C, Schindler A. Reliability and validity of the Italian version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). Folia Phoniatr Logop. 2013;65(5):257–65.
10 de Krom G. Consistency and reliability of voice quality ratings for different types of speech fragments. J Speech Hear Res. 1994 Oct;37(5):985–1000.
11 Revis J, Giovanni A, Wuyts F, Triglia J. Comparison of different voice samples for perceptual analysis. Folia Phoniatr Logop. 1999; 51(3):108–16.
12 Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. Folia Phoniatr Logop. 2002 Nov-Dec;54(6):271–81.
13 Law T, Kim JH, Lee KY, Tang EC, Lam JH, van Hasselt AC, et al. Comparison of Rater's reliability on perceptual evaluation of different types of voice sample. J Voice. 2012 Sep; 26(5):666.e13–21.
14 Fex B, Fex S, Shiromoto O, Hirano M. Acoustic analysis of functional dysphonia: before and after voice therapy (accent method). J Voice. 1994 Jun;8(2):163–7.
15 Lechien JR, Finck C, Khalife M, Huet K, Delvaux V, Picalugga M, et al. Change of signs, symptoms and voice quality evaluations throughout a 3- to 6-month empirical treatment for laryngopharyngeal reflux disease. Clin Otolaryngol. 2018 May. https://doi.org/10.1111/coa.13140.
16 Finck C. Implantation d'acide hyaluronique estérifié lors de la microchirurgie des lésions cordales bénignes [PhD thesis]. Liege: University of Liege; 2008.
17 Lu FL, Matteson S. Speech tasks and interrater reliability in perceptual voice evaluation. J Voice. 2014 Nov;28(6):725–32.
18 Zraick RI, Wendel K, Smith-Olinde L. The effect of speaking task on perceptual judgment of the severity of dysphonic voice. J Voice. 2005 Dec;19(4):574–81.

19 Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. J Speech Hear Res. 1993 Feb;36(1):14–20.
20 Maryn Y, Roy N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. J Soc Bras Fonoaudiol. 2012;24(2):107–12.
21 Bele IV. Reliability in perceptual analysis of voice quality. J Voice. 2005 Dec;19(4):555–73.
22 Soni RS, Ebersole B, Jamal N. Does Even Low-Grade Dysphonia Warrant Voice Center Referral? J Voice. 2017 Nov;31(6):753–56.
23 Ghio A, Dufour S, Wengler A, Pouchoulin G, Revis J, Giovanni A. Perceptual evaluation of dysphonic voices: can a training protocol lead to the development of perceptual categories? J Voice. 2015 May;29(3):304–11.
24 Belafsky PC, Postma GN, Koufman JA. Laryngopharyngeal reflux symptoms improve before changes in physical findings. Laryngoscope. 2001 Jun;111(6):979–81.
25 Freitas FF, Costa KN, Rebouças CB, Fernandes M, Lima JO. [Nonverbal communication between nurses and the elderly based on the proxemics]. Rev Bras Enferm. 2014 Nov-Dec;67(6):928–35.
26 Eadie TL, Kapsner M, Rosenzweig J, Waugh P, Hillel A, Merati A. The role of experience on judgments of dysphonia. J Voice. 2010 Sep;24(5):564–73.
27 Wolfe V, Fitch J, Martin D. Acoustic measures of dysphonic severity across and within voice types. Folia Phoniatr Logop. 1997;49(6):292–9.
28 Choi SH, Lee J, Sprecher AJ, Jiang JJ. The effect of segment selection on acoustic analysis. J Voice. 2012 Jan;26(1):1–7.
29 Núñez Batalla F, Corte Santos P, Sequeiros Santiago G, Señaris González B, Suárez Nieto C. [Perceptual evaluation of dysphonia: correlation with acoustic parameters and reliability]. Acta Otorrinolaringol Esp. 2004 Jun-Jul;55(6):282–7.
30 Kelchner LN, Brehm SB, Weinrich B, Middendorf J, deAlarcon A, Levin L, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. J Voice. 2010 Jul;24(4):441–9.
31 Anders LC, Hollien H, Hurme P, Sonninen A, Wendler J. Perception of hoarseness by several classes of listeners. Folia Phoniatr (Basel). 1988;40(2):91–100.
32 Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. J Acoust Soc Am. 2000 Oct;108(4):1867–76.
33 Gerratt BR, Kreiman J, Garellek M. Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech. J Speech Lang Hear Res. 2016 Oct;59(5):994–1001.
34 Schoentgen J, Fraj S, Lucero JC. Testing the reliability of Grade, Roughness and Breathiness scores by means of synthetic speech stimuli. Logoped Phoniatr Vocol. 2015 Apr;40(1):5–13.
35 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. J Voice. 1997 Mar;11(1):74–80.
36 Moon KR, Chung SM, Park HS, Kim HS. Materials of acoustic analysis: sustained vowel versus sentence. J Voice. 2012 Sep;26(5):563–5.
37 Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. Folia Phoniatr Logop. 2009;61(1):49–56.