Observer Agreement for Measurements in Videolaryngostroboscopy

Observer Agreement for Measurements in Videolaryngostroboscopy

*,†,‡Jan Wouter Brunings, §Sophie Vanbelle, *Annemarie Akkermans, *Nienke M.M. Heemskerk, *,†Bernd Kremer, *,‡Robert J. Stokroos, and *,†Laura W.J. Baijens, *†‡§*Maastricht, The Netherlands*

Summary: Objective. This study evaluated the levels of intraobserver and interobserver agreement for measurements of visuoperceptual variables in videolaryngostroboscopic examinations and compared the observers' behavior during independent versus consensus panel rating.

Study Design. This is a retrospective study.

Setting. This study was conducted in a single-center tertiary care facility.

Participants. Sixty-four patients with dysphonia of heterogeneous etiology were included.

Exposure. All subjects underwent a standardized videolaryngostroboscopic examination.

Main Outcome and Measures. Two experienced and trained observers scored exactly the same examinations, first independently and then on a consensus panel. Specific visuoperceptual variables and the clinical diagnosis (as recommended by the Committee on Phoniatrics and the Phonosurgery Committee of the European Laryngological Society and advised by the American Speech-Language-Hearing Association) were scored. Descriptive and kappa statistics were used.

Results. In general, intraobserver agreement was better than agreement between observers for measurements of several variables. The intrapanel observer agreement levels were slightly higher than the intraobserver agreement levels on the independent rating task. When rating on the consensus panel, the observers deviated considerably from the scores they had previously given on the independent rating task.

Conclusion and Relevance. Observer agreement in videolaryngostroboscopic assessment has important implications not only for the diagnosis and treatment of dysphonic patients but also for the interpretation of the results of scientific studies using videolaryngostroboscopic outcome parameters. The identification of factors that can influence the levels of observer agreement can provide a better understanding of the rating process and its limitations. The results of this study suggest that future research could achieve better agreement levels by rating the visuoperceptual variables in a panel setting.

Key Words: Videolaryngostroboscopy–Voice–Observer agreement–Reliability–Visuoperceptual variables.

INTRODUCTION

Laryngoscopy was first used to examine the larynx and the vocal folds at the end of the 19th century, when Manuel Garcia visualized the vocal folds.[1] Since then, the technique has been refined.[2] Now, videolaryngostroboscopy is an important tool for the clinical assessment of vocal pathologies and function, and it can be used to evaluate the effectiveness of treatments.[3] Clinicians around the world apply videolaryngostroboscopy in daily clinical practice, using many variables to describe vocal fold functioning and the vibratory pattern, although data on the reliability and validity of these variables remain scarce. Several studies report the levels of intraobserver and interobserver agreement for measurements in videolaryngostroboscopy, but few mention which visuoperceptual variables are used in daily clinical practice. Thus, it is unknown whether those studies used the visuoperceptual variables recommended by the Committee on Phoniatrics of the European Laryngological Society (ELS)[3] and advised by the American Speech-Language-Hearing Association (ASHA),[4] most of which are based on the parameters defined by Hirano and Bless.[5]

External validation of clinical decisions based on the interpretation of videolaryngostroboscopic exams requires accurate and reliable measurements.[6] The description of aspects that can influence observer behavior and agreement on measured videolaryngostroboscopic variables can elucidate the rating process and thereby facilitate developing procedures to increase the observer agreement levels.

This study has two aims: (1) to evaluate the intraobserver, interobserver, and intrapanel agreement for measurements in videolaryngostroboscopic examinations and (2) to compare the observers' behavior during independent versus consensus panel rating. It was hypothesized that intrapanel agreement would sustain higher levels of observer reproducibility.

MATERIAL AND METHODS

Selection of participants

Videolaryngostroboscopic recordings of 64 patients with dysphonia of heterogeneous etiology were selected by an independent coworker laryngologist from the clinical research archives of the Maastricht University Medical Center. All patients were referred to the multidisciplinary outpatient clinic for their dysphonia

Accepted for publication September 6, 2017.

From the *Department of Otorhinolaryngology, Head and Neck Surgery, Maastricht University Medical Center, Maastricht, The Netherlands; †GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands; ‡MHeNs-School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, The Netherlands; and the §Department of Methodology and Statistics, CAPHRI-School for Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands.

Address correspondence and reprint requests to Jan Wouter Brunings, Department of Otorhinolaryngology, Head and Neck Surgery, Maastricht University Medical Center, P.O. Box 5800, Maastricht 6202 AZ, The Netherlands. E-mail: jw.brunings@mumc.nl

Journal of Voice, Vol. 32, No. 6, pp. 756–762
0892-1997

https://doi.org/10.1016/j.jvoice.2017.09.005

and underwent a standardized videolaryngostroboscopic examination as part of the regular health-care program. During the selection process, videolaryngostroboscopic recordings were excluded if the patient presented severe discomfort or gagging. Reasons for exclusion were disturbing body movements and a diagnosis of head and neck cancer. Only recordings made at least 1 year previously to the study in 2014 were included to avoid recognition of patients by the laryngologists and to ensure blinding to patient identity.

Videolaryngostroboscopic examination

All videolaryngostroboscopic examinations and measurements were performed in the same hospital by the same two laryngologists (each with over 10 years of experience in performing and rating videolaryngostroboscopy). All subjects underwent a standardized videolaryngostroboscopic examination. The videos were recorded using a 70- or 90-degree rigid Hopkins endoscope (model 8706 CA; Karl Storz GmbH & Co KG, Tuttlingen, Germany), which was attached to an Alphatron Stroboview ACLS camera, Alphatron Lightsource, IVACX computerized video archiving system (Alphatron Medical Systems Rotterdam, the Netherlands), and were recorded on a DVD. A rigid endoscopy was carried out to obtain an image of a quality superior to that attainable with flexible equipment. If necessary, a topical anesthetic (Xylocaine 10% (AstraZeneca AB, Södertälje, Sweden)) was applied. The examination had to be stable enough to allow continuous stroboscopic "tracking" of vocal fold vibration. Care was taken to ensure that each selected video contained a phonation time long enough to allow for the registration of a sustained phonation of the vowel /i:/, at comfortable pitch and loudness, and at least one complete cycle of vibration. The audio track was removed from the recordings. During the examination, subjects were seated upright wearing their dental prosthesis if present. The field of the image included the laryngeal vestibule, vocal folds, anterior commissure, and arytenoids. Informed consent was obtained from all subjects.

Study design and measurements

Prior to data collection, the two laryngologists received training in visuoperceptual measurement for the nominal and ordinal variables given in Table 1. The categorical scales of the variables recommended by the ELS and ASHA were modified as described in Table 1. The videos used for training were not included in the experimental set of videos. Training in the exact interpretation of each scale category took place in sessions with both observers and was intended to generate substantial to almost perfect levels of intraobserver and interobserver agreement (kappa 0.61–0.99). The duration of the program was predetermined. It consisted of two training sessions, approximately 1 hour each, interspersed over the course of a month, with practice periods when the observers had to do test runs separately. A written manual with well-defined descriptions of the scales' levels was available during the training and the rating process for the observers to consult. Visuoperceptual nominal and ordinal variables and the clinical diagnosis (nominal variable) were scored for each videolaryngostroboscopic video at varying speeds if desired by the rater (slow motion, normal, up to frame-by-frame) as often as necessary using the software program *Windows Movie Maker* version 5.1 (Microsoft Corporation, Redmond, WA). Clinical diagnosis of the vocal pathology was derived from reports of the Phonosurgery Committee of the ELS (Table 2).[8]

The rating process comprised two separate tasks: independent rating and consensus panel rating of the same randomized videolaryngostroboscopic videos. While the two observers were blinded to each other's ratings during the independent task, the decision on the score was reached in consensus during the panel

TABLE 1.
Visuoperceptual Nominal and Ordinal Variables Used to Rate Videolaryngostroboscopic Recordings

Variable Name	Definition	Scale
Diagnosis of vocal pathology*	Diagnosis derived from reports of the Phonosurgery Committee of the European Laryngological Society[7]	Not applicable (see Table 2)
Amplitude, left and right vocal folds†	Extent of vocal fold displacement near the glottic opening	0 = normal 1 = impaired 2 = absent
Periodicity, left and right vocal folds†	Temporal regularity of vibratory cycles	0 = normal 1 = impaired
Symmetry†	Symmetry of mucosal displacement	0 = normal 1 = impaired
Closure†	Degree of glottic closure during the closed phase of vibration	0 = normal 1 = impaired
Defect*	Type of glottic closure: predominant mucosal closure patterns	0 = normal 1 = oval 2 = hourglass 3 = anterior 4 = posterior (<50%) 5 = complete (>50%)

* Visuoperceptual nominal variable.
† Visuoperceptual ordinal variable.

TABLE 2.
Clinical Diagnosis of Vocal Pathology, as Assessed by the Panel, Derived from Reports of the Phonosurgery Committee of the European Laryngological Society, in Absolute Numbers (N) and Percentages (%)

Diagnosis	N (%)
Anterior webbing	2 (3.2)
Atrophy	3 (4.7)
Granuloma vocal process	5 (7.8)
Intracordal hematoma	1 (1.6)
Intracordal cyst	8 (12.5)
Laryngitis	10 (15.6)
Laryngitis sicca	1 (1.6)
Laryngocele	1 (1.6)
Nodules	1 (1.6)
Normal larynx	9 (14.1)
Polyp	0 (0.0)
Papillomatosis	0 (0.0)
Paralysis	11 (17.2)
Reinke space edema III/IV	3 (4.7)
Reinke space edema I/II	7 (10.9)
Scarification	2 (3.1)

task, which took place 1 month later. Differences between the raters during the panel task were solved by discussion using the manual with well-defined descriptions of the scales' levels and controlled mutual feedback.

The observers were blinded to the vocal sound recording and to the medical history and identity of the patients.

To determine how well the intraobserver scores arrived at jointly agreed with those obtained independently, the laryngologists (again blinded) repeated the assessments within a period of 4 months for all variables on 32 videos randomly selected from the 64 initial videos. They were advised to limit the duration of the measurement sessions (maximum 1 hour) to avoid fatigue and prevent instability of observers' input.

Statistical analysis

Results were expressed as means (standard deviation) for quantitative variables, while frequencies and proportions (%) were used for categorical variables. For the nominal variables, ie, periodicity, symmetry, closure, diagnosis, and defect, disagreement on a video was defined as a difference in the classification made by the two observers. For ordinal variables, ie, amplitude, disagreement was defined as the number of categories separating the classification made by the two observers. In this latter case, a disagreement on two adjacent categories of the same measurement scale is therefore considered less important than a disagreement on more distant categories. The percentage of videos on which two observers agree was reported. However, agreement between two observers can also occur even if they randomly allocate the videos. Consequently, the level of agreement was adjusted for the amount of agreement expected by chance, leading to Cohen's kappa coefficient for nominal variables and the linear weighted kappa coefficient for ordinal variables. The intraobserver and interobserver agreement level, quantified through kappa coefficients, was reported with 95% confidence interval (CI).

TABLE 3.
Demographic and Clinical Characteristics of the 64 Dysphonic Patients and the Frequency Distribution of Patients per Category of the Videolaryngostroboscopic Variables in Absolute Numbers (N) and Percentages (%)

Variable		N (%)	Mean (SD)
Gender	Man	30 (50.0)	
	Woman	30 (50.0)	
Age	Total		50.3 (17.8)
	Man		56.6 (17.3)
	Woman		44.3 (16.4)
Amplitude—left side	Normal	18 (34.6)	
	Impaired	28 (53.8)	
	Absent	6 (11.5)	
Amplitude—right side	Normal	26 (44.2)	
	Impaired	26 (50.0)	
	Absent	3 (5.8)	
Periodicity—left side	Normal	33 (71.7)	
	Impaired	13 (28.3)	
Periodicity—right side	Normal	36 (78.2)	
	Impaired	10 (21.7)	
Symmetry	Normal	8 (14.0)	
	Impaired	49 (86.0)	
Closure	Normal	8 (13.1)	
	Impaired	53 (86.9)	
Defect	Normal	8 (13.1)	
	Oval	12 (19.7)	
	Hourglass	9 (14.8)	
	Anterior	3 (4.9)	
	Posterior	16 (26.2)	
	Complete	13 (21.3)	

The standard error of the kappa coefficient used to construct the 95% CI was adjusted for the presence of repeated measurements when necessary, ie, when a variable was assessed on both sides of the larynx of the same patient.[9] Not accounting for the repeated measurements will artificially increase the sample size and lead to CIs that are too narrow.

A comparison of the intraobserver agreement levels obtained by the two observers individually and jointly was made using Hotelling's T squared statistic, which is a generalization of Student' t test to more than two variables. The variance-covariance matrix used in the calculations was estimated using a bootstrap method.[10,11] Missing data were not replaced, ie, no imputation. Data analysis was performed using *R* (version 3.0.1 for Windows [Microsoft Corporation]).

RESULTS

Characteristics of the participants

The demographics of the dysphonic patients and their clinical characteristics, as assessed by the panel, are displayed in Table 3.

TABLE 4.
Interobserver Agreement Levels for the Two Independent Observers on the 64 Dysphonic Patients

	N	% Agreement	Kappa*	SE	95% CI
Diagnosis	64	67.1	0.63	0.064	0.50–0.75
Amplitude	101	57.4	0.37	0.090	0.19–0.55
Left	50	62.0	0.46	0.10	0.14–0.77
Right	51	52.9	0.28	0.11	–0.15–0.57
Periodicity	94	71.0	0.30	0.11	0.084–0.52
Left	46	72.0	0.36	0.12	0.12–0.61
Right	48	71.0	0.22	0.11	0.0021–0.43
Symmetry	57	86.0	0	NA	NA
Closure	63	98.0	0.85	0.15	0.56–1.00
Defect	54	70.4	0.62	0.077	0.47–0.77

<0, less than chance agreement.
0.01–0.20, slight agreement.
0.21– 0.40, fair agreement.
0.41–0.60, moderate agreement.
0.61–0.80, substantial agreement.
0.81–0.99, almost perfect agreement.
* Kappa coefficient.
Abbreviations: NA, not available; SE, standard error.

The frequency distribution of patients per category of the videolaryngostroboscopic variables is shown in Table 3, giving an indication of the average voice function of the study population.

Patients (30 men and 30 women, four missing values) were, on average, 50.3 years old, the men being older than the women (mean: 56.6 versus 44.3 years, $P = 0.007$). The study population was a fair reflection of the dysphonic patients who consult our outpatient clinic. The most frequently observed conditions, as assessed by the panel, were paralysis (N = 11; 17.2%) laryngitis (N = 10; 15.6%), normal larynx (N = 9; 14.1%), intracordal cyst (N = 8; 12.5%), and Reinke edema grades I and II (N = 7; 10.9%) (Table 2).

Agreement study

Interobserver agreement levels are given in Table 4. Levels of intraobserver agreement for the two individual observers and the panel are shown in Table 5. Table 6 gives the number of changes in the video scores made by the observers during the panel meeting compared with their individual assessment of exactly the same videos. Amplitude and periodicity were measured separately for the left and right sides. For these variables, agreement is determined for each side separately and at the patient level.

Diagnosis

Interobserver agreement for clinical diagnosis was 0.63 (95% CI: 0.50–0.75) (Table 4). Intraobserver agreement was 0.67 (0.50–0.84) for observer 1, 0.75 (0.59–0.91) for observer 2, and 0.78 (0.62–0.93) for both during the panel rating task (Table 5). No particular disagreement pattern was observed. Nor was any significant difference in intraobserver agreement found between the panel and the individual assessments ($P = 0.57$). However, the clinical diagnosis based on the exact same videos during the panel rating task and the individual assessments changed for 27 (25.8%) patients (Table 6).

Amplitude

Interobserver agreement for the variable amplitude was 0.37 (0.19–0.55) (Table 4). Disagreement arose on the 61 videos classified as "impaired" by the first observer. Among these videos, 14 (23%) were classified as "normal" and 12 (19.7%) as "absent amplitude" by the second observer. Intraobserver agreement was 0.42 (0.18–0.67) for observer 1, 0.37 (0.15–0.59) for observer 2, and 0.76 (0.53–0.92) for the observers together during the panel rating task (Table 5). Here too, more than 50% of the videos initially classified as "impaired" were subsequently classified as "normal" or "absent amplitude" by the team. Intraobserver agreement was higher during the panel meeting than during individual assessments ($P = 0.016$) (Table 5). After discussion, the observers modified the amplitude classification of 61 (30.3%) videos, most of which had been classified as "impaired" or "absent" during their individual assessments (Table 6).

Periodicity

Interobserver agreement for the variable periodicity was 0.30 (0.084–0.52) (Table 4). While the first observer classified 36 (38.3%) videos as impaired, only 11 (11.7%) were classified thus by the second observer. Intraobserver agreement was 0.51 (0.17–0.86) for observer 1, 1 (not applicable) for observer 2, and 0.37 (–0.023–0.76) for the observers together (Table 5). No particular disagreement pattern was observed. Forty-nine (27.2%) of the videos scored individually by the observers were rescored at the panel meeting (Table 6).

Symmetry

Interobserver agreement for the variable symmetry was 0, although 86% of the videos were classified in the same category by the two observers (Table 4). In fact, all videos were classified as "impaired" by the first observer, explaining the difference between the high percentage of agreement and the low value of

TABLE 5.
Intraobserver Agreement Levels for Each Observer and for the Consensus Panel on 32 Dysphonic Patients

	Observer 1			Observer 2			Panel		
	N	% Agreement	Kappa* (SE)	N	% Agreement	Kappa (SE)	N	% Agreement	Kappa (SE)	*P* Value
Diagnosis	32	71.9	0.67 (0.086)	32	78.1	0.75 (0.081)	32	81.3	0.78 (0.079)	0.57
Amplitude	48	58.3	0.42 (0.13)	47	55.3	0.37 (0.17)	52	78.8	0.76 (0.17)	0.016
Left	24	58.3	0.45 (0.24)	23	56.5	0.38 (0.25)	32	76.9	0.71 (0.25)	0.053
Right	24	54.2	0.36 (0.22)	24	54.2	0.37 (0.24)	26	80.8	0.74 (0.24)	0.008
Periodicity	46	76.1	0.51 (0.18)	46	100	1 (NA)	43	72.1	0.37 (0.20)	0.50†
Left	23	78.3	0.56 (0.29)	23	100	1 (NA)	22	68.2	0.34 (0.27)	0.32†
Right	23	73.9	0.46 (0.27)	23	100	1 (NA)	21	76.2	0.39 (0.21)	0.76†
Symmetry	29	93.1	–0.036 (0.025)	26	53.8	0.20 (0.10)	26	92.3	0.70 (0.20)	0.0010
Closure	30	96.7	0.78 (0.21)	31	96.7	0.78 (0.21)	30	83.3	0.19 (0.24)	0.012
Defect	24	75.0	0.59 (0.25)	31	54.8	0.46 (0.24)	30	60.0	0.52 (0.25)	0.24

<0, less than chance agreement.
0.01–0.20, slight agreement.
0.21– 0.40, fair agreement.
0.41–0.60, moderate agreement.
0.61–0.80, substantial agreement.
0.81–0.99, almost perfect agreement.
* Kappa coefficient.
† Comparing only the intraobserver agreement for observer 1 and the panel.
Abbreviations: NA, not available; SE, standard error.

Cohen's kappa coefficient. Among these, 49 (86%) were classified as impaired by the second observer. Intraobserver agreement was –0.036 (–0.085; 0.014) for observer 1, 0.20 (0.0040–0.40) for observer 2, and 0.70 (0.32–1.00) for the panel (Table 5). This result is in contrast with the percentage of agreement between the two measurement occasions (93.1% for observer 1, 53.8% for observer 2, and 92.3% for the panel) (Table 5). Here too, the difference between the percentage of agreement and the kappa coefficient is explained by the low number of patients classified with normal symmetry. Normal symmetry was observed only in 1 (3.4%) patient for observer 1, 10 (38%) patients for observer 2, and 4 (15.4%) patients for the panel rating task. Intraobserver agreement was higher for the panel task than for the independent assessments ($P = 0.0010$) (Table 5). The individually assigned classifications of 22 (20.8%) videos were changed during the panel meeting (Table 6).

Closure

Interobserver agreement for closure was 0.85 (0.56–1.00) (Table 4). Intraobserver agreement was 0.78 (0.02–1.00) for observer 1, 0.78 (0.37–1.00) for observer 2, and 0.19 (–0.28–0.67) for the observers jointly (Table 5). No particular pattern of disagreement was seen on this variable. Normal closure was rarely noted (maximum two [6.7%] videos by observer 1, three [10%] for observer 2, and four [13.3%] for the panel). Intraobserver agreement was low for the panel, even though 25 (83.3%) videos were classified in the same category by the panel on both measurement occasions. Intraobserver agreement was higher for individual assessments than for the panel ($P = 0.012$) (Table 5). During the panel meeting, 11 (9.1%) videos were reclassified (Table 6).

Defect

Interobserver agreement for the variable defect was 0.62 (0.47–0.77) (Table 4). Intraobserver agreement was 0.59 (0.47–0.90) for observer 1, 0.46 (0.24–0.67) for observer 2, and 0.52 (0.28–0.72) for the panel (Table 5). No particular disagreement pattern was seen. There was no significant difference between the individual and consensus panels for intraobserver agreement ($P = 0.24$) (Table 5). The classification of 51 (44.7%) videos was changed during the panel meeting (Table 6).

DISCUSSION

Visuoperceptual analysis of videolaryngostroboscopic images has some limitations and is subject to error. However, no other alternative is available to analyze these images in daily clinical practice. While recourse to videolaryngostroboscopy is increasing, research on the standardization and validation of measurement criteria in these exams lags behind. Crucially, the interpretation of videolaryngostroboscopic images rests on visual judgment and is thus subjective.[6] It might be influenced by factors such as experience, observer fatigue, severity of vocal fold disease, etc.[12] Investigating observer agreement is the first step to demonstrating the validity of the procedure because measurement criteria can only be valid if they are reproducible in terms of observer agreement.[13] Some studies have addressed observer agreement on some well-known visuoperceptual

TABLE 6.
Scores Given Independently Compared with Scores Given on the Consensus Panel for Exactly the Same Measurement. Number (%) of Changes in the Scores Arising When Comparing the Individual to Consensus Measurements*

	Observer 1 Independently; First Measurement	Observer 1 Independently; Repeated Measurement	Observer 2 Independently; First Measurement	Observer 2 Independently; Repeated Measurement	Total
	Number (%) of Changes	Number (%) of Changes	Number (%) of Changes	Number (%) of Changes	Number (%) of Changes
Diagnosis	11 (34.4)	2 (6.3)	6 (18.8)	8 (25.0)	27 (25.8)†
Amplitude	17 (34.0)	10 (20.0)	18 (35.0)	16 (32.0)	61 (30.3)
Left	7 (28.0)	5 (20.0)	8 (32.0)	10 (40.0)	30 (30.0)
Right	10 (40.0)	5 (20.0)	10 (38.5)	6 (24.0)	31 (30.7)
Periodicity	16 (39.0)	9 (18.7)	6 (14.6)	18 (36.0)	49 (27.2)
Left	8 (38.0)	5 (20.8)	3 (14.3)	10 (40.0)	26 (28.6)
Right	8 (40.0)	4 (16.7)	3 (15.0)	8 (32.0)	23 (25.8)
Symmetry	5 (17.9)	3 (12.0)	2 (7.1)	12 (48.0)	22 (20.8)
Closure	4 (12.9)	1 (3.4)	4 (12.9)	2 (6.7)	11 (9.1)
Defect	13 (52.0)	6 (21.4)	19 (61.3)	13 (43.3)	51 (44.7)

* To facilitate the interpretation of the table, a more detailed description is given here. For instance, in the row "Diagnosis," in 11 videos observer 1 changed his initial score (given during the independent rating task) when he rated these same video recordings in the panel setting. Observer 1 rated 35 video recordings independently a second time to obtain his intraobserver agreement level. On these previously rated video recordings, he changed his score twice when he rated the same videos in the panel setting.
† This column represents the summed score of all the changes made in the initial scores after discussion by the two observers in the panel setting.

variables.[12] Nonetheless, the variability in the scoring of videolaryngostroboscopic exams remains underexplored. Given its role in clinical decision making, such as on (surgical) treatment, these exams must be both accurate and reliable.

The present study considers intraobserver and interobserver agreement for visuoperceptual measurements in videolaryngostroboscopic exams and explores any discrepancies that arise between independent and consensus ratings to better understand the causes of disagreement among observers. In this study, observer agreement was determined for specific visuoperceptual variables and clinical diagnoses that are used in daily clinical practice derived from the recommendations of the Committee on Phoniatrics and the Phonosurgery Committee of the ELS and advised by the ASHA. For the independent rating task, intraobserver agreement on several variables (diagnosis, amplitude, and closure) was similar for both observers. This suggests that the two observers had a similar interpretation of the nominal and ordinal scoring systems and were consistent when repeating the measurements. In general, intraobserver agreement was better than agreement between different observers for measurements of almost all variables except for "closure" and "defect." The finding that intraobserver agreement was better than agreement between observers for the majority of variables might be explained as follows. First, even though the observers understood the nominal and ordinal scoring systems well, they did not reach consensus on the cutoff points. The description of the rating scale does not give the precise range of each nominal or ordinal category, which leaves it up to the observers to set their own boundaries. Accordingly, the interpretation of videolaryngostroboscopic variables is often subject to a wide degree of variability because of the subjective discrimination of the levels of these scales.

In an overview of the literature, studies investigating the intraobserver and interobserver agreement of measurements in videolaryngostroboscopic exams were found to have methodological shortcomings. Notably, interobserver agreement varied considerably with the method of measurement, the level of experience of the observers, the number and pathology of the subjects, and the presence or absence of the audio track (auditive information).[7,14–20] Often, it was not reported whether the observers had received pre-experimental training.[20–22] In general, the conclusions could not be compared across the studies because of diverging study designs, small populations, and different ways of evaluating the videolaryngostroboscopic exams. Given their heterogeneous methodologies, the studies could not be compared with one another. Some had issues concerning the generalization of results. For instance, performing several measurements on a large number of participants or healthy subjects in a limited time period makes the exercise susceptible to bias and, moreover, does not reflect clinical practice.[15,18,20] Therefore, observer agreement varies across raters and populations.[7,12,16,17,20,23,24]

In the present study, intrapanel observer agreement was slightly higher than the intraobserver levels on the independent rating task (except for the variables periodicity and closure). That difference suggests that consensus rating might be a useful alternative to the independent rating of videolaryngostroboscopic exams, as the discussion of cases in a panel may improve concordance. However, the level of agreement between two separate consensus panels with different members still needs to be explored, particularly in comparison with individual interobserver

agreement levels. Also, the validity of the measurement criteria of the different variables should be investigated.

Observers were consistent for several measured variables (diagnosis, periodicity, closure, defect) when rescoring videos independently or on the consensus panel. However, when repeating the task on the panel, they frequently adjusted the scores they had given previously when rating exactly the same measurements independently. That tendency to change in a panel setting reflects the observers' individual interpretation of the videolaryngostroboscopic scoring system.

Limitations of the study

The results of the study might be affected by the limited number of videos analyzed. However, considering the amount of time observers spent during the pre-experimental training period and the rating process, a larger number of videos would have made the process last so long that it would be difficult to engage experienced observers. This study presents the videolaryngostroboscopic scores of just two observers. Comparing scores of a larger number of observers could yield different results. However, the situation in which the present study was conducted resembles most clinical settings where one or two professionals are usually responsible for the interpretation of videolaryngostroboscopic exams.

CONCLUSION

Observer agreement for videolaryngostroboscopic measurements has important implications not only for the diagnosis and treatment of dysphonic patients but also for the interpretation of the results of studies using videolaryngostroboscopic outcome parameters. The identification of factors that can influence the observer agreement levels such as the setting of the rating task (independent versus panel) can provide a better understanding of the rating process and its limitations. For research purposes, this study suggests that the visuoperceptual variables should be rated in a panel to achieve better agreement levels. The same procedure is suggested for the clinical setting: to use an experienced panel composed, for example, of the laryngologist and speech and language pathologist who work together at the outpatient voice clinic.

REFERENCES

1. Garcia M. Physiological observations on the human voice. *Proc R Soc Lond*. 1855;7:339.
2. Heman-Ackah YD. Diagnostic tools in laryngology. *Curr Opin Otolaryngol Head Neck Surg*. 2004;12:549–552.
3. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
4. Association AS-L-H. Vocal tract visualization and imaging: technical report, 2004.
5. Hirano M, Bless DM. *Videostroboscopic Examination of the Larynx*. San Diego, CA: Singular Publishing Group Inc.; 1993.
6. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-based clinical voice assessment: a systematic review. *Am J Speech Lang Pathol*. 2013;22:212–226.
7. Nospes S, Kuhr K, Napiontek U, et al. Stroboscopy findings: a comparison of flexible CCD-videostroboscopy and rigid stroboscopy. *Laryngorhinootologie*. 2011;90:218–223.
8. Friedrich G, Remacle M, Birchall M, et al. Defining phonosurgery: a proposal for classification and nomenclature by the Phonosurgery Committee of the European Laryngological Society (ELS). *Eur Arch Otorhinolaryngol*. 2007;264:1191–1200.
9. Yang Z, Zhou M. Kappa statistic for clustered matched-pair data. *Stat Med*. 2014;33:2612–2633.
10. Vanbelle S, Albert A. A bootstrap method for comparing correlated kappa coefficients. *J Stat Comput Sim*. 2008;78:1009–1015.
11. Vanbelle S. Comparing dependent kappa coefficients obtained on multilevel data. *Biom J*. 2017;59:1016–1034.
12. Rosen CA. Stroboscopy as a research instrument: development of a perceptual evaluation tool. *Laryngoscope*. 2005;115:423–428.
13. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59:1033–1039.
14. Gelfer MP, Bultemeyer D. Evaluation of vocal fold vibratory patterns in normal voices. *J Voice*. 1990;4:335–345.
15. Kendall KA. High-speed laryngeal imaging compared with videostroboscopy in healthy subjects. *Arch Otolaryngol Head Neck Surg*. 2009;135:274–281.
16. Nawka T, Konerding U. The interrater reliability of stroboscopy evaluations. *J Voice*. 2012;26:812, e1–10.
17. Poburka BJ. A new stroboscopy rating form. *J Voice*. 1999;13:403–413.
18. Pemberton C, Russell A, Priestley J, et al. Characteristics of normal larynges under flexible fiberscopic and stroboscopic examination: an Australian perspective. *J Voice*. 1993;7:382–389.
19. Dejonckere PH, Crevier L, Elbaz E, et al. Quantitative rating of videolaryngostroboscopy: a reliability study. *Rev Laryngol Otol Rhinol (Bord)*. 1998;119:259–260.
20. Lundy DS, Casiano RR, Sullivan PA, et al. Incidence of abnormal laryngeal findings in asymptomatic singing students. *Otolaryngol Head Neck Surg*. 1999;121:69–77.
21. Teitler N. Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. *J Voice*. 1995;9:95–105.
22. Olthoff A, Woywod C, Kruse E. Stroboscopy versus high-speed glottography: a comparative study. *Laryngoscope*. 2007;117:1123–1126.
23. Chau HN, Desai K, Georgalas C, et al. Variability in nomenclature of benign laryngeal pathology based on video laryngoscopy with and without stroboscopy. *Clin Otolaryngol*. 2005;30:424–427.
24. Cipriani NA, Martin DE, Corey JP, et al. The clinicopathologic spectrum of benign mass lesions of the vocal fold due to vocal abuse. *Int J Surg Pathol*. 2011;19:583–587.