



<b>Title</b>	<b>Reliability of laryngo-stroboscopic evaluation based on visual perceptual judgment</b>
<b>Author(s)</b>	Lau, Chi-yan; 劉智恩
<b>Citation</b>	Lau, C. [劉智恩]. (2013). Reliability of laryngo-stroboscopic evaluation based on visual perceptual judgment. (Thesis). University of Hong Kong, Pokfulam, Hong Kong SAR.
<b>Issued Date</b>	2013
<b>URL</b>	<a href="http://hdl.handle.net/10722/238523">http://hdl.handle.net/10722/238523</a>
<b>Rights</b>	This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.; The author retains all proprietary rights, (such as patent rights) and the right to use in future works.

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

**Reliability of laryngo-stroboscopic evaluation based on  
visual perceptual judgment**

Lau Chi Yan

A dissertation submitted in partial fulfilment of the requirements for the Bachelor of Science  
(Speech and Hearing Sciences), The University of Hong Kong, June 30, 2013.

**Abstract**

This study investigated the inter-rater and intra-rater reliability of visual perceptual evaluation of laryngo-stroboscopic images. Two hundreds and fifty-five laryngo-stroboscopic videos samples were collected from 75 subjects. Three raters undertook evaluation of the images on 4 measurements: 1) mass lesion size, 2) amplitude of vocal fold vibration, 3) supraglottic activity, and 4) shape of the glottal closure, using the modified Stroboscopy Examination Rating Form (Poburka, 1999). Results showed that substantial inter-rater and intra-rater reliability were achieved for lesion size, antero-posterior supraglottic activity and glottal closure. However, evaluation of medio-lateral supraglottic activity and the amplitude of vocal fold vibration could not achieve an adequate reliability (ranged from 0.45-0.50). The finding indicated that laryngo-stroboscopic examination is a relatively reliable method for the measurement of lesion size, antero-posterior supraglottic compression and glottal closure. This finding is better than those reported in the literature (Nawka & Konerding, 2012). Meanwhile, the vocal fold vibratory amplitude measure was found to be the least reliable.

**Reliability of laryngo-stroboscopic evaluation based on  
visual perceptual judgment**

Laryngo-stroboscopy has been used as a clinical tool for the diagnosis of voice disorders since the late 1800s (Colton, Casper, & Leonard, 2006). It allows visualization of the vibratory patterns of vocal folds, which is difficult to be observed with human eyes due to the high vibrational frequency during phonation. Laryngo-stroboscopy creates an optical illusion of slow motion that enables clinicians to visualize the vibration of vocal folds by using a series of images captured over a number of successive vocal cycles. It is widely agreed as an important technique for laryngeal examination as it provides invaluable information for the evaluation of voice disorders (Poburka & Bless, 1998; Rosen, 2005). Cohen, Pitman, Noordzij, and Courey (2012) found that 84% of the 273 responded general otolaryngologists from the American Academy of Otolaryngology – Head and Neck Surgery performed laryngo-stroboscopic examination in routine practice. In addition, laryngo-stroboscopy is also frequently used as a research tool or an evaluation tool in studies which investigate the efficacy of different treatment approaches for voice disorder (Lorenz et al., 2008; Speyer et al., 2002; Wang et al., 2011).

The clinical value of laryngo-stroboscopy has been supported by a number of studies. Woo, Colton, Casper, and Brewer (1991) carried out 195 laryngo-stroboscopic examinations on 146 patients to study the importance of laryngo-stroboscopy. Laryngo-stroboscopy was

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

found to contribute critical diagnostic information in 27.2% of the examinations, which means there was alternation of the diagnosis made without the use of laryngo-stroboscopy or it could give diagnosis that could not be made before having laryngo-stroboscopic examination. Moreover, laryngo-stroboscopy provided supplementary information for the original diagnosis in 48.7% of the examinations. In another study which evaluated 292 patients, Casiano, Zaveri, and Lundy (1992) found 19% of the patients, who were originally evaluated by indirect laryngoscopy, required a change in their diagnoses after performing laryngo-stroboscopic examination. Additional findings and change in treatment methods were made in 24% and 11% of them respectively. Laryngo-stroboscopic examination was also found to have significant improvement on the assessment of glottis closure. Hernández Sandemtrio, Nieto Curiel, Dalmau Galofre, and Forcada Barona (2010) compared the diagnosis obtained by laryngo-stroboscopy preoperatively and the intraoperative diagnosis in 91 patients with a total of 181 lesions. Correlation between the two diagnoses was observed in 90% of the cases for oedema, polyp, nodule, intracordal cyst and fibrosis.

### **Limitation of laryngo-stroboscopic examination**

As laryngo-stroboscopic examination mainly relies on visual perceptual judgments, its subjective nature raises the concern of the reliability of the measurements made (Nawka & Konerding, 2012; Teitler, 1995). Reliability of a measurement refers to how well it can differentiate between subjects or items (Kottner et al., 2011). Hirano and Bless (1993)

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

identified several factors that might affect the interpretation of laryngo-stroboscopic images, which included the knowledge of the observers on the anatomy and physiology of vocal fold vibration, the skills with the laryngo-stroboscopic technique and skills in interpreting the images. Another possible bias in laryngo-stroboscopic evaluation includes prior knowledge of the patient's case history (Teitler, 1995).

### **Statistics for reliability**

In order to understand the reliability figures reported in the literature, a review of the statistics of reliability will be provided here. There are a number of indices that can be used to indicate the reliability of a measurement. Commonly used indices include intraclass correlation coefficient (ICC) and kappa statistics. Percent agreement, which is an index for agreement, is also sometimes reported in studies evaluating research instruments.

For interval and ratio data, ICC is the most widely used index in the literature. It is described as the most appropriate index for the measurement of reliability (Tinsley & Brown, 2000). ICC is an index of the proportion of the target variance to the total variance. It is used to evaluate the concordance of the measurements made by two raters or more, and can apply to studies with missing ratings (Gisev, Bell, & Chen, in press). Different models of ICC have been designed for different situation and purposes. According to Shrout and Fleiss (1979), one-way random effect model is used when the targets are rated by a different group of raters. However, if all the targets are rated by the same group of raters, a two-way

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

model would be used. Depending on whether the effects of raters or repeated trials are assumed to be random, the two-way model can be further classified as two-way random effects model and two-way mixed model (Shrout & Fleiss, 1979). For each of the models, there are two specifications of ICC: ICC for absolute agreement and ICC for consistency.

The ICC for consistency only considers the general ranking of the ratings given by the raters, while the ICC for absolute agreement considers the absolute value of the ratings in addition to the ranking. Therefore, when the ratings given by a rater are consistently higher than those given by another rater, the value of ICC for consistency will be higher than that for absolute agreement (Gisev et al., in press).

Kappa is another index that has been commonly reported in the literature to assess reliability or agreement (Gisev et al., in press). It is a chance-corrected index suitable for nominal and ordinal scales (Sim & Wright, 2005). There are three assumptions for kappa: 1) the items to be rated are independent, 2) the raters are independent, and 3) the categories of the scale are independent (Tinsley & Brown, 2000). Cohen's kappa is a type of kappa statistics used to assess the inter-rater and intra-rater reliability when the items are rated by two raters and when two ratings are given by one rater respectively (Sim & Wright, 2005). For studies which involve more than two raters, Fleiss' kappa, which is an extension of kappa, would be a possible option. Siegel and Castellan (1988) also discussed a version of multi-rater kappa statistic which can be used to assess reliability with three or more raters.

## Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

Percent agreement is a simple index which can give a general idea of the degree of agreement in a study (Lombard, Snyder-Duch, & Bracken, 2002). It is calculated as the number of ratings in agreement divided by the total number of ratings. While reliability index provides information of how well an instrument can differentiate between individuals, percent agreement tells us the degree to which the ratings are identical (Gisev et al., in press). Although percent agreement is easy to calculate and can be used in any number of raters, it does not account for the agreement achieved by chance. Despite this limitation, Kottner et al. (2011) recommended to report percent agreement in combination with other indices so as to give an overall impression of the extent of reliability and agreement to the readers.

### **Reliability reported for laryngo-stroboscopic evaluation**

As laryngo-stroboscopic evaluation is a subjective process, a number of rating methods have been developed to quantify laryngo-stroboscopic evaluation in order to achieve satisfactory reliability (Poburka, 1999; Rosen, 2005). Studies that examined the inter-rater and intra-rater reliability of laryngo-stroboscopic evaluation have reported variable reliability figures. Nawka and Konerding (2012) examined 68 video clips to evaluate the inter-rater reliability of laryngo-stroboscopy. They found that the intraclass correlation coefficient (ICC) ranging from 0.32-0.71. Using the definition given by Landis and Koch (1977), only four of the 15 measures demonstrated adequate reliability (0.55-0.60) or substantial reliability (0.61-0.80). Phase closure, phase symmetry and regularity received the lowest reliability



Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

(0.34, 0.41 and 0.32 respectively). Nawka and Konerding (2012) argued that laryngo-stroboscopy only showed the average vocal fold movement and thus, short-term changes of movements, which were important in evaluating phase closure, phase symmetry and regularity, would be difficult to observe especially when the changes were small. Hence, laryngo-stroboscopy would not be an appropriate technique to evaluate these measures.

In another study with 21 unique video samples (Rosen, 2005), the interclass correlation value for inter-rater reliability reported ranged from 0.11-0.68. Intra-rater reliability as determined using Kendall's Tau- $\beta$  was found to range from 0.43-0.99. It should be noted that only four out of the 10 measures received adequate or substantial reliability. In addition, a contrasting result was obtained for the reliability of amplitude of left vocal fold (0.61) and amplitude of right vocal fold (0.25).

Poburka (1999) evaluated the reliability of laryngo-stroboscopic examination using 42 video samples. The results for percentage of exact inter-rater agreement varied greatly from 2%-71%, with a majority of the measures received a percentage of lower than 50%. Nevertheless, better results were found for the percentage of consensus which were ranged from 53%-100%.

Among the studies that evaluated the reliability of laryngo-stroboscopic evaluation, the lesion size on vocal fold is seldom included as one of the measures for investigation. The size of lesion in laryngo-stroboscopic examination is important as changes in size reflect

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

change in condition over time or as a consequence of treatment. Kobler et al. (2006)

contended that accurate description of the laryngeal lesion size would allow more precise

staging and grading of the pathologies, which could lead to better treatment decisions and

outcomes. Hence, a reliable rating scale for lesion size is essential (Shah, Feldman, & Nuss,

2007). Shah et al. (2007) developed a four-point rating scale for pediatric vocal fold

nodules based on static images, and the ICC for inter-rater reliability ranged from 0.67-0.87.

As noted that laryngeal examination was mainly based on the review of video clips clinically,

but not static images, Nuss, Ward, Recko, Huang, and Woodnorth (2012) further validated the

rating form developed by Shah et al. (2007) using video clips. Inter-rater reliability

calculated using ICC ranged from 0.52-0.74 with a mean of 0.62, and the overall ICC for

intra-rater reliability was 0.86.

### **Aim of the Study**

In the study by Nawka and Konerding (2012), after excluding the measures that were not

suitable for rating using laryngo-stroboscopy (i.e. phase closure, phase symmetry and

regularity), the reliability for amplitude (0.44), supraglottic activity (0.42) and glottal closure

(0.38) were among the lowest. As these three measures are important clinical measures and

are commonly evaluated in laryngo-stroboscopic assessment, their reliability was therefore

re-examined in the present study to determine if similar reliability were to be found.

In addition, most studies usually reported only one rating for supraglottic activity when

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

investigating its reliability. Indeed, supraglottic activity can be observed in two directions, medio-lateral (M-L) and antero-posterior (A-P). A-P compression occurs when the arytenoid cartilages moves towards the petiole of the epiglottis, while M-L compression is caused by the adduction of the ventricular folds. The Stroboscopy Examination Rating Form (Poburka, 1999) is one of the evaluation forms that provides rating of supraglottic activity in both the M-L and A-P directions. Nevertheless, Poburka (1999) reported inter-rater percentages of exact agreement and consensus, which did not sufficiently reflect how the ratings made by the raters deviated from each other. The present study, therefore, examined the reliability of supraglottic activity in both the M-L and A-P directions using intraclass correlation coefficient (ICC) in addition to the agreement.

Accurate description of the lesion size will facilitate better documentation, treatment decision and outcomes. Shah et al. (2007) and Nuss et al. (2012) investigated the reliability for the rating of nodule size in the pediatric population using a categorical scale based on the extent of nodule protrusion (normal, nodule protrudes  $< 0.5$  mm, between 0.5-1.0 mm or  $>1.0$  mm). However, this rating scale might not be applicable to adult case as the size of vocal folds in adults is much larger. Thus, lesion size was selected as one of the measures so as to assess its reliability in adult cases. The present study used the grid concept proposed by Poburka (personal communication, see Appendix A) and the rating was made by estimating the number of grids covered by the mass lesion.

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

It was noted that most of the studies reviewed above only had around 20-40 video samples for ratings, which might lead to limited variability in some of the measures (Poburka, 1999; Poburka & Bless, 1998; Rosen, 2005). As the variability of a specific measure in the samples would affect the reliability of the evaluation (Murphy & Davidshofer, 2005), the present study used a larger number of video samples (a total 225 video samples), so as to achieve a wide variability for the reliability to be assessed.

Therefore, the present study examined the inter-rater and intra-rater reliability of laryngo-stroboscopic evaluation based on visual perceptual judgment. It aimed to determine how reliable laryngo-stroboscopic examination was in different measures. Four sets of measures were examined in the present study: vocal fold vibratory amplitude (left and right), supraglottic activity (antero-posterior and medio-lateral), glottal closure and the lesion size on vocal fold (left and right). Therefore, a total of seven measures were evaluated.

## **Method**

### **Participants**

The data in the present study were collected from two earlier studies which investigated the effect of acupuncture treatment (K. L. Ho, 2012; L. K. Ho, 2012). Laryngo-stroboscopic video samples were collected from 75 subjects who participated in the studies mentioned above. They aged between 20 and 56 years old with a mean age of 39.21 and standard deviation of 10.60. They were diagnosed as having dysphonia associated with benign

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

pathological tissues changes. The diagnoses for the pathologies included nodules (N=43), polyps (N=14), thickened vocal folds (N=11), chronic laryngitis (N=6), gap on adduction (N=1) and granuloma (N=1). Some of the subjects had two pathologies. Three raters, each with more than 10 years of experience in laryngo-stroboscopic examination, took part in rating the video samples independently.

### **Materials**

Laryngo-stroboscopic examinations using rigid endoscopy (Storz Model Xenon 300, Storz telecam Model SLII and Kay Rhino-Laryngeal Stroboscope Model 9100B) were performed on the 75 subjects at the Voice Research Laboratory of the University of Hong Kong. Corel Video Studio Pro X2 image processing software (Corel, Corel Video Studio 12) was used to record and edit the images. Each subject had three sets of recording, one before treatment, one immediately after the treatment, and the third one recorded one month after the completion of treatment. Video clips from 10 subjects were duplicated for rating in order to determine intra-rater reliability. Hence, a total of 255 laryngo-stroboscopic video samples ( $75 \times 3 + 10 \times 3$ ) were obtained. Each video recording was approximately 5-10 seconds in duration and was recorded under quiet respiration and sustained phonation of /i/ at a comfortable pitch and loudness level.

### **Procedures**

The order of presentation of the video clips within each subject set was randomized and

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

the order of presentation of the subject sets was also randomized. All videos were presented using a 68.6 cm LED monitor with a resolution of 2560 x 1440 (Apple LED display). An anchor laryngoscopic image obtained from a female with a normal voice and complete glottal closure was placed next to each video to be rated so as to provide a common reference point for the raters. The raters were blind to the diagnosis and treatment stage of the subjects in the videos. The raters could control the speed of presentation themselves and were allowed to view the video repeatedly if needed. A modified version of the Stroboscopy Examination Rating Form (SERF) (Poburka, 1999) (Appendix A) was used. For the rating of amplitude, the original five-point rating scale was used. The ratings ranged from small vibratory movement (20%) to full vibratory movement (100%) with an interval of 20%. The raters had to give rating for the left and right vocal folds separately. Two supraglottic activity measures were rated (medio-lateral (M-L) and antero-posterior (A-P)) using a six-point scale, where 0 represented no compression and 5 represented maximum compression to the midline. Glottal closure was categorized into complete closure, anterior gap, posterior gap, hourglass shape, spindle gap shape, irregular and incomplete closure. For the rating of the size of the vocal lesion, the raters were required to sketch the lesion on the form and then count the number of grids the lesion occupies on both sides of the vocal folds (see Appendix A).

### **Statistical analysis**

The data were analyzed with the Statistical Package for the Social Sciences 17.0 (IBM

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

SPSS Inc, Chicago, IL). Inter-rater and intra-rater reliability for the interval measures (i.e. amplitude, supraglottic activity and nodule size) was assessed using ICC. As generalization of the result to other studies and clinical setting was intended in this study, the effect of raters (for inter-rater reliability) and repeated trials (for intra-rater reliability) were considered as random effect for the calculation of ICC. Hence, the two-way random effect model with the specification of ICC for absolute agreement was used. Intra-rater reliability for the categorical measure (i.e. glottal closure) was calculated using Cohen's kappa as the repeated sets of subjects were rated twice. Since the SPSS procedure only offers Cohen's kappa for two raters, a macro for the version of multi-kappa statistic discussed by Siegel and Castellan (1988) was downloaded from the statistical macro library of the IBM to evaluate the inter-rater reliability in this study (Appendix B) (Nichols, 1997). In addition, agreements were also calculated as this would provide a different perspective of the reliability (Kottner et al., 2011). Exact agreement (i.e. for both interval and categorical data) and agreement within one point of scale (i.e. for interval data only) between the three raters were reported.

## **Results**

### **Inter-rater reliability**

The inter-rater reliability calculated using ICC and kappa for the seven measures is summarized in Table 1. The reliability values ranged from 0.45 to 0.81. Four of the measures attained substantial reliability (0.61-0.80) as defined by Landis and Koch (1977).

## Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

They were left and right lesion size, AP supraglottic activity and glottal closure. The reliability measures for the two amplitude measures and ML supraglottic activity were all smaller than 0.55, which is the lower limit for adequate reliability (0.55-0.60) according to Landis and Koch (1977). The two amplitude measures had the lowest reliability among all the measures. The reliability of the AP supraglottic activity (0.64) was found to be higher than that of the ML supraglottic activity ML (0.50).

Table 1. *Inter-rater reliability of different measures*

Parameter	Inter-rater reliability	<i>p</i> value	Limit of 95% interval
Lesion size			
Right	(ICC) 0.81	< .001	0.77-0.85
Left	(ICC) 0.75	< .001	0.69-0.79
Amplitude			
Right	(ICC) 0.46	< .001	0.38-0.54
Left	(ICC) 0.45	< .001	0.37-0.53
Supraglottic activity			
Antero-posterior	(ICC) 0.64	< .001	0.58-0.70
Medio-lateral	(ICC) 0.50	< .001	0.43-0.58
Glottal closure	(Kappa) 0.65	< .001	(0.60-0.71)

Table 2 shows the exact agreement and agreement within one point measure for the three



## Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

raters. As the glottal closure is a categorical parameter, only exact agreement was calculated. For any two raters, all the exact agreement measures were larger than 0.50 except for the AP supraglottic activity between raters 1 and 3 (0.47). The agreement within one point measure were greater than 0.90.

For the agreement among the three raters, the exact agreement measures were ranged from 0.33 to 0.68, while the agreement within one point measures were all higher than 0.89. The supraglottic activity measures showed the lowest agreement.

Table 2. *Inter-rater agreement of different measures*

Parameter	Raters 1 & 2		Raters 1 & 3		Raters 2 & 3		All raters	
	Exact	+/- 1	Exact	+/- 1	Exact	+/- 1	Exact	+/- 1
Lesion size								
Right	0.58	0.92	0.59	0.96	0.62	0.96	0.44	0.89
Left	0.60	0.93	0.58	0.93	0.66	0.96	0.46	0.88
Amplitude								
Right	0.79	1.00	0.76	0.99	0.79	1.00	0.68	0.99
Left	0.68	1.00	0.70	1.00	0.80	1.00	0.59	0.99
Supraglottic activity								
Antero-posterior	0.52	0.95	0.47	0.94	0.60	0.97	0.33	0.89
Medio-lateral	0.52	0.96	0.56	0.97	0.64	0.98	0.39	0.92

## Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

Glottal closure	0.72	0.80	0.75	0.64
-----------------	------	------	------	------

---

+/- 1: Agreement for ratings within one point measure

It should be noted that although the amplitude measures showed the lowest reliability values in ICC, the agreement values for the amplitude ratings were generally higher than those for the other measures. In contrast, the lesion size rating showed the highest ICC values among other measures, demonstrated relatively lower exact agreement when compared with the other measures such as glottal closure and amplitude.

### **Intra-rater reliability**

The intra-rater reliability of the three raters for all the measures was listed in Table 3. The inter-rater reliability for the lesion size ratings (0.65-0.95, substantial reliability according to Landis and Koch (1977) ) were generally higher than those for the other measures (0.20-0.70). The intra-rater reliability for supraglottic activity and glottal closure was similar across the raters and they were ranged from 0.39 to 0.56 and 0.57 to 0.65 respectively. The intra-rater reliability for the amplitude measures showed a greater variability across different raters.

Table 3. *Intra-rater reliability of the three raters on different measures*

Parameter	Rater 1	Rater 2	Rater 3
	Reliability <i>p</i> value	Reliability <i>p</i> value	Reliability <i>p</i> value
Lesion size			

## Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

Right	(ICC)	0.95	< .001	0.76	< .001	0.91	< .001
Left	(ICC)	0.73	< .001	0.65	< .001	0.86	< .001
Amplitude							
Right	(ICC)	0.20	.146	0.70	< .001	0.63	< .001
Left	(ICC)	0.34	.035	0.60	< .001	0.23	.104
Supraglottic activity							
Antero-posterior	(ICC)	0.56	< .001	0.55	< .001	0.54	.001
Medio-lateral	(ICC)	0.39	.014	0.39	.017	0.54	.001
Glottal closure	(Kappa)	0.65	< .001	0.60	< .001	0.57	< .001

Table 4 summarized the intra-rater exact agreement and agreement within one point measure for the three raters. As the glottal closure is a categorical data, only exact agreement data were calculated. The exact agreement measures were all approximately 0.60 or above in rating the lesion size and amplitude of vibration. The AP and ML supraglottic activities, however, demonstrated exact agreement from 0.27 to 0.80. The agreement within one point measures was all 0.85 or higher.

Table 4. *Intra-rater agreement of the three raters on different measures*

Parameter	Raters 1		Raters 2		Raters 3	
	Exact	+/- 1	Exact	+/- 1	Exact	+/- 1

Lesion size

## Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

Right	0.80	0.97	0.60	0.90	0.70	0.97
Left	0.57	0.97	0.67	0.87	0.73	1.00
Amplitude						
Right	0.63	1.00	0.83	1.00	0.83	1.00
Left	0.67	1.00	0.77	1.00	0.67	1.00
Supraglottic activity						
Antero-posterior	0.27	0.90	0.43	0.93	0.33	0.97
Medio-lateral	0.37	0.97	0.57	0.97	0.80	1.00
Glottal closure	0.77		0.70		0.70	

---

+/- 1: Agreement for ratings within one point measure

When comparing between the reliability and agreement data, contrasting results were obtained in the amplitude and supraglottic activity ratings. While the ICC values for amplitude varied highly (0.20-0.70), the exact agreement measures for amplitude were relatively higher and less variable across the three raters (0.63-0.83). On the other hand, the reliability data for the supraglottic activity ratings were less variable across the three raters (0.39-0.55) than the exact agreement data (0.27-0.80).

### Discussion

The objective of the present study was to investigate the reliability of four measures used in laryngo-stroboscopic examination. They were lesion size, amplitude, supraglottic

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

activity and glottal closure. These measures were found to have low reliability (i.e. amplitude, supraglottic activity and glottal closure) or seldom investigated (i.e. lesion size) in previous studies. Therefore, the reliability in using these laryngo-stroboscopic measures was re-examined in the present study.

### **Lesion size**

The result for the reliability of measuring lesion size (left and right) was encouraging since substantial reliability calculated using ICC was attained for both the inter-rater and intra-rater reliability in all the raters. Satisfactory agreement was obtained for lesion size with an average of 0.60 inter-rater and intra-rater exact agreement. There were very few studies reported in the literature that investigated the reliability of evaluating vocal fold mass lesion size in adults. The positive result from the present study provided empirical support for the use of laryngo-stroboscopy as a reliable tool for evaluating vocal fold lesion size in adults (c.f. Shah et al., 2007 & Nuss et al, 2012).

### **Amplitude of vocal fold vibration**

Vocal fold vibratory amplitude rating was found to have inadequate reliability. It received the lowest inter-rater and intra-rater reliability measures in general. Although a relatively good agreement was obtained for amplitude, it was probably due to the narrow range of ratings (low variability) given by the three raters. As more than 97% of the videos were given an amplitude rating of 20% or 40%, the chance level of agreement between the

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

raters on these two amplitude ratings would be high. This also explained why 1.00 agreement within one point could be easily achieved in inter-rater agreement. Thus, the agreement measure was not sufficient to reflect the reliability of the amplitude measure in laryngo-stroboscopic examination. The ICC values found in the present study for amplitude rating were consistent with those found by Nawka and Konerding (2012), therefore suggesting that the reliability for amplitude rating was not satisfactory in laryngo-stroboscopic examination based entirely on visual perceptual judgment.

As the variability of ratings in a sample would affect the reliability of a measurement (Murphy & Davidshofer, 2005), the low reliability obtained for amplitude might be attributed to the low variability present in the subjects. To assure adequate variability, a review on the video samples by another group of raters can be done before the rating procedure. Rosen (2005) also suggested checking the variability of the videos samples in a consensus format before the study in order to enhance the reliability. Another possible explanation for the low variation in the ratings was that the scale for amplitude might not be sensitive enough to differentiate between patients and thus, the ratings for amplitude were highly concentrated in 20% and 40%. Therefore, further investigation was recommended to determine if modifying the scale for amplitude would help improve the reliability.

### **Supraglottic activity**

It should be noted that the reliability for measuring AP supraglottic activity was

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

generally higher than that for the ML supraglottic activity in both inter-rater and intra-rater reliability. The agreement within one point for both measures was about 0.90 or higher in all cases. As the endoscope might not be placed vertically above the glottis during each laryngo-stroboscopic examination, there were some variations in the viewing angles in the video samples. Hirano and Bless (1993) pointed out that image might be distorted due to variation in the angle of view and the distance between the endoscope tip and the object. Kobler et al. (2006) also stated that the measurement error caused by varying viewing angle could be affected by the size and shape of a structure. Since the antero-posterior dimension of the supraglottic activity is greater than the medio-lateral dimension, it was suspected that the difference in viewing angle might affect the observation of the supraglottic activity in the medio-lateral directions more than that in the antero-posterior direction. Therefore, it might be more difficult for the raters to make judgment on the supraglottic activity in the medio-lateral directions and hence, led to a lower reliability.

Previous studies had found that supraglottic activity in the medio-lateral direction was also present in normal voiced individuals and there was no significant difference between normal individuals and individuals with voice disorder (Behrman, Dahl, Abramson, & Schutte, 2003; Stager et al., 2001). On the other hand, significant difference in antero-posterior compression was found between normal and voice-disordered groups. Therefore, the measurement of supraglottic activity in the antero-posterior direction might

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

have a greater diagnostic value for voice disorder than that in medio-lateral direction.

Hence, although the reliability for ML supraglottic activity was not satisfactory, measurement of AP supraglottic activity was, however, relatively reliable.

### **Glottal closure**

The reliability for identifying the types of glottal closure was also encouraging in both inter-rater and intra-rater measures. The results of the present study on identifying glottal closure type contradicted with those of Nawka and Konerding (2012), who found a low reliability using the SERF from Poburka (1999) as in the present study. They even suggested that glottal closure pattern evaluation should not be included in laryngo-stroboscopic procedure because of its low reliability. A possible reason for the difference in findings was that the larger number of videos used in the present study provided greater variability, which allowed reliability to be better reflected.

In a visual perceptual evaluation task, the quality of the images is of utmost importance. If the images are not clear enough, it would be difficult to rate the samples and thus, the process would be highly un-reliable. Therefore, control on the quality of the samples should not be neglected in study investigating reliability of a measurement.

Poburka and Bless (1998) proposed that the measures evaluated in laryngo-stroboscopic examination could be classified into two categories, which were geometric and dynamic



Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

measures. Geometric measures required raters to make judgment on the shape or configuration of a structure while dynamic measures involved judgment of a continuous movement pattern. Geometric measures were argued to be easier to rate as raters were only required to make simple description on the physical appearance. Dynamic measures, which involved evaluation a continuously changing movement, were relatively difficult for ratings. Since lesion size and glottal closure could be classified as geometric measures, they received generally lower reliability than the amplitude and supraglottic activity, which were dynamic in nature. Hence, the intrinsic nature of the amplitude and supraglottic activity measures, might have contributed to the poor reliability obtained in laryngo-stroboscopic examination.

In the present study, ICC values or Kappa coefficient were reported in conjunction with the agreement data. Our results showed that a high agreement in the ratings does not necessarily reveal good reliability when compared with the ICC or kappa coefficients.

Although agreement did not take into account the agreement that are expected by chance, it did provide a general picture of how the raters varied in their ratings as observed in the study.

### **Limitation**

As mentioned before, the videos samples used in the study were collected from subjects participated in two treatment studies and three videos were obtained from each subject.

Although the videos from each subject were presented in a random order, they were arranged together and the raters would know which videos were belonged to the same subject.

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

Therefore, it might affect the judgment made by the raters in some extent due to their awareness of the close relatedness of the video sets. Since the three videos obtained from each subject were treated as three individual samples in the study, this effect was not considered when evaluating the reliability of the measures.

In addition, the poor quality of the video samples as reported by the raters was another limitation since it could undoubtedly affect the reliability obtained in the study. To avoid the same problem in future study, a selection of videos should be done to ensure the videos used for investigated are of good quality.

### **Conclusion**

In the present study, the reliability of four laryngo-stroboscopic measures was investigated. The result revealed that good reliability was obtained for the measurement of lesion size and glottal closure. Hence, it supported the use of laryngo-stroboscopy as a reliable instrument to evaluate these measures. Although the reliability for supraglottic activity in the medio-lateral direction was not adequate, the reliability for the antero-posterior direction, which has a more important diagnostic value, was found to be satisfactory. The vibrational amplitude of the vocal fold had the lowest reliability and it might be due to the low variability in the subjects or the small scale interval. Hence, further investigation may be carried out to examine if increasing the variability in the subjects or modification of the rating scale for amplitude will help improve its reliability. .

### References

- Behrman, A., Dahl, L. D., Abramson, A. L., & Schutte, H. K. (2003). Anterior-posterior and medial compression of the supraglottis: signs of nonorganic dysphonia or normal postures? *Journal of Voice*, *17*(3), 403-410.
- Casiano, R. R., Zaveri, V., & Lundy, D. S. (1992). Efficacy of videostroboscopy in the diagnosis of voice disorders. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, *107*(1), 95-100.
- Cohen, S. M., Pitman, M. J., Noordzij, J. P., & Courey, M. (2012). Evaluation of dysphonic patients by general otolaryngologists. *Journal of Voice*, *26*(6), 772-778.
- Colton, R. H., Casper, J. K., & Leonard, R. L. (2006). *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Baltimore, MD: Lippincott Williams & Wilkins.
- Gisev, N., Bell, J. S., & Chen, T. F. (in press). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*.
- Hernández Sandemetro, R., Nieto Curiel, P., Dalmau Galofre, J., & Forcada Barona, M. (2010). What is the contribution of stroboscopy in the diagnosis of voice disorders? *Acta Otorrinolaringologica (English Edition)*, *61*(2), 145-148.
- Hirano, M., & Bless, D. M. (1993). *Videostroboscopic examination of the larynx*. San Diego,

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

Calif: Singular Publishing Group Inc.

Ho, K. L. (2012). *Effectiveness of laser acupuncture for hyperfunctional dysphonia - a treatment placebo group study*. (Unpublished bachelor dissertation), The University of Hong Kong, Hong Kong.

Ho, L. K. (2012). *Effect of acupuncture on the healing of benign vocal fold lesions*. (Unpublished bachelor dissertation), The University of Hong Kong, Hong Kong.

Kobler, J. B., Rosen, D. I., Burns, J. A., Akst, L. M., Broadhurst, M. S., Zeitels, S. M., & Hillman, R. E. (2006). Comparison of a flexible laryngoscope with calibrated sizing function to intraoperative measurements. *Annals of Otolaryngology, Rhinology & Laryngology*, 115(10), 733-740.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., . . . Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies*, 48(6), 661-671.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4), 587-604.

Lorenz, R. R., Esclamado, R. M., Teker, A. M., Strome, M., Scharpf, J., Hicks, D., . . . Lee, W.

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

- T. (2008). Ansa cervicalis-to-recurrent laryngeal nerve anastomosis for unilateral vocal fold paralysis: Experience of a single institution. *Annals of Otolology, Rhinology & Laryngology*, *117*(1), 40-45.
- Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications*. Upper Saddle River, N. J.: Perason/Prentice Hall.
- Nawka, T., & Konerding, U. (2012). The interrater reliability of stroboscopy evaluations. *Journal of Voice*, *26*(6), 812.e811-812.e810.
- Nichols, D. (1997). MKAPPASC.SPS. Retrieved March 31, 2013, from <ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Macros/mkappasc.sps>
- Nuss, R. C., Ward, J., Recko, T., Huang, L., & Woodnorth, G. H. (2012). Validation of a pediatric vocal fold nodule rating scale based on digital video images. *Annals of Otolology, Rhinology & Laryngology*, *121*(1), 1-6.
- Poburka, B. J. (1999). A new stroboscopy rating form. *Journal of Voice*, *13*(3), 403-413.
- Poburka, B. J., & Bless, D. M. (1998). A multi-media, computer-based method for stroboscopy rating training. *Journal of Voice*, *12*(4), 513-526.
- Rosen, C. A. (2005). Stroboscopy as a research instrument: Development of a perceptual evaluation tool. *The Laryngoscope*, *115*, 423-428.
- Shah, R. K., Feldman, H. A., & Nuss, R. C. (2007). A grading scale for pediatric vocal fold

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

nodules. *Otolaryngology - Head and Neck Surgery*, 136(2), 193-197.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability.

*Psychological Bulletin*, 86(2), 420-428.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*.

New York: McGraw-Hill.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation,

and sample size requirements. *Physical Therapy*, 85(3), 257-268.

Speyer, R., Kempen, P. A., Wieneke, G., Kersing, W., Hosseini, E. G., & Dejonckere, P. H.

(2002). Effects of voice therapy as objectively evaluated by digitized laryngeal

stroboscopic imaging. *Annals of Otology, Rhinology & Laryngology*, 111(10), 902.

Stager, S. V., Bielamowicz, S., Gupta, A., Marullo, S., Regnell, J. R., & Barkmeier, J. (2001).

Quantification of static and dynamic supraglottic activity. *Journal of Speech,*

*Language, and Hearing Research*, 44(6), 1245-1256.

Teitler, N. (1995). Examiner bias: Influence of patient history on perceptual ratings of

videostroboscopy. *Journal of Voice*, 9(1), 95-105.

Tinsley, H. E. A., & Brown, S. D. (2000). *Handbook of applied multivariate statistics and*

*mathematical modeling*. San Diego: Academic Press.

Wang, W., Chen, D., Chen, S., Li, D., Li, M., Xia, S., & Zheng, H. (2011). Laryngeal

reinnervation using ansa cervicalis for thyroid surgery-related unilateral vocal fold

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

paralysis: A long-term outcome analysis of 237 cases. *PLoS one*, 6(4), 1-7.

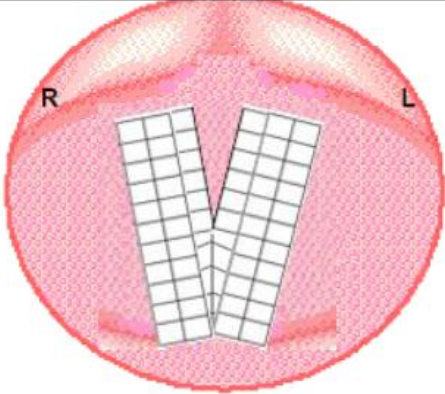
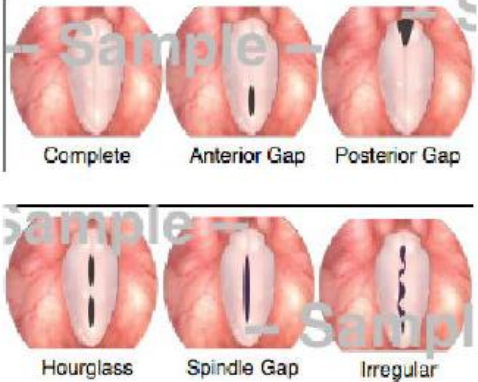

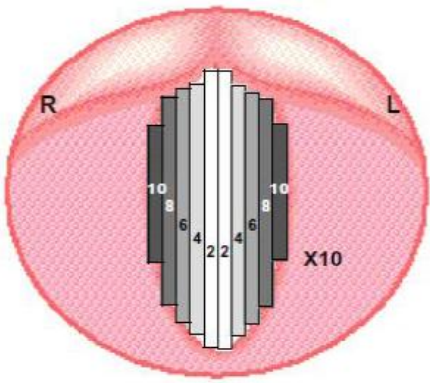
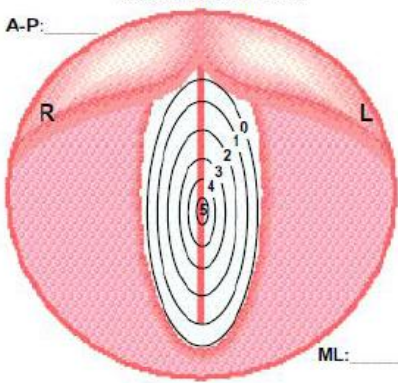
Woo, P., Colton, R., Casper, J., & Brewer, D. (1991). Diagnostic value of stroboscopic

examination in hoarse patients. *Journal of Voice*, 5(3), 231-238.

**Appendix A**

A modified version of the Stroboscopy Examination Rating Form (SERF) based on Poburka

(1999)

<p>Please mark the extent of lesion on the grid</p>  <p><b>Right:</b> _____ Number of boxes: _____</p> <p><b>Left:</b> _____ Number of boxes: _____</p>	<p>Please circle the glottal closure</p>  <p><b>Glottal closure:</b> Appearance of glottis during the most closed portion of the glottal cycle</p>  <p>Incomplete</p>
<p>Please mark the extent of vibratory movement from midline according to the scale (2,4,6,8,10)</p> <p><b>Amplitude</b> (Rate @ normal pitch &amp; loudness)</p>  <p><b>Right:</b> _____%      <b>Left:</b> _____%</p> <p><b>Amplitude</b> Extent of vibratory movement of vocal folds laterally to midline.</p>	<p>Please mark the degree of compression according to the scale (0-5)</p> <p><b>Supraglottic Activity</b> (Ignore voice onsets)</p> <p>A-P: _____</p>  <p><b>Supraglottic activity</b> <i>Medio-lateral compression (ML)</i> true vocal folds being obscured by the ventricular folds. <i>Antero-posterior compression (A-P)</i> compression or shortening of the aryepiglottic folds.</p>



**Appendix B**

Macro for the calculation of the version of multi-rater kappa statistic

preserve.

set printback=off mprint=off.

save outfile='ka\_\_tmp1.sav'.

define mkappasc (vars=!charend('/')).

set mxloops=1000.

count ms\_\_=!vars (missing).

select if ms\_\_=0.

matrix.

get x /var=!vars.

compute c=mmax(x).

compute y=make(nrow(x),c,0).

loop i=1 to nrow(x).

loop j=1 to ncol(x).

loop k=1 to c.

do if x(i,j)=k.

compute y(i,k)=y(i,k)+1.

end if.

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

end loop.

end loop.

end loop.

compute pe=msum((csum(y)/msum(y))&\*\*2).

compute k=ncol(x).

compute pa=mssq(y)/(nrow(y)\*k\*(k-1))-(1/(k-1)).

compute kstat=(pa-pe)/(1-pe).

compute num=2\*(pe-(2\*k-3)\*(pe\*\*2)+2\*(k-2)\*msum((csum(y)/msum(y))  
&\*\*3)).

compute den=nrow(y)\*k\*(k-1)\*((1-pe)\*\*2).

compute ase=sqrt(num/den).

compute z=kstat/ase.

compute sig=1-chicdf(z\*\*2,1).

save {kstat,ase,z,sig} /outfile='ka\_\_tmp2.sav'

    /variables=kstat,ase,z,sig.

end matrix.

get file='ka\_\_tmp2.sav'.

formats all (f11.8).

variable labels kstat 'Kappa' /ase 'ASE' /z 'Z-Value' /sig 'P-Value'.

Running head: RELIABILITY OF LARYNGO-STROBOSCOPIC EVALUATION

report format=list automatic align(center)

/variables=kstat ase z sig

/title "Estimated Kappa, Asymptotic Standard Error,"

"and Test of Null Hypothesis of 0 Population Value".

get file='ka\_\_tmp1.sav'.

!enddefine.

restore.

mkappasc vars= rater1 to rater3.

### **Acknowledgements**

I would like to sincerely thank my supervisor, Prof. Edwin Yiu, for his unconditional guidance and support throughout the study. I would also like to thank Dr. Estella Ma and Dr. Karen Chan for their precious comments and generous assistance in rating the laryngoscopic videos. Special thanks go to Gladys and Masato for providing me with the data collected in their studies.