View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HKU Scholars Hub

Title	**Interrater and intrarater reliability in rating velopharyngeal gap size**
Other Contributor(s)	**University of Hong Kong.**
Author(s)	**Leung, Hei-man, Heman**
Citation	
Issued Date	**2007**
URL	**http://hdl.handle.net/10722/55481**
Rights	**The author retains all proprietary rights, such as patent rights and the right to use in future works.**

Interrater and intrarater reliability in rating velopharyngeal gap size

Leung, Hei Man Heman

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, 30th June, 2007

Interrater and intrarater reliability in rating velopharyngeal gap size

Leung Hei Man, Heman

Abstract

The use of nasendoscopy as an assessment tool for evaluation of velopharyngeal function has been widely advocated. The interpretation of assessment results remains perceptual in nature. Golding-Kushner et al. (1990) developed a standardized protocol for the reporting of nasendoscopy findings. This standard has been widely accepted. However, research to assess the reliability of the scale is still limited. Nasendoscopy assessment serves important clinical and research purposes, for example, decision making on the need for and type of secondary surgery. Therefore, the reliability of the assessment is an important issue. In addition, factors which might affect the reliability are unknown. This study has two research aims: The first was to investigate interrater and intrarater reliability in rating velopharyngeal gap size; the second was to investigate if factors such as speech sample, velopharyngeal configuration and quality of nasendoscopy recording affected reliability. Three expert raters were asked to rate velopharyngeal gap size using a 6-point scale adapted from Golding-Kushner et al. (1990). The results revealed satisfactory correlation between (0.76, $p < 0.05$) raters. Fair to good correlations were found within raters (0.42 - 0.74, $p < 0.05$). However, no significant findings were obtained concerning the possible factors which might affect reliability. It was concluded that this 6-point scale was reliable for rating velopharyngeal gap size for clinical and research purposes. Further research should focus on understanding the potential effect of experience on the reliability of rating and the effectiveness of training for interpreting nasendoscopy results.

INTRODUCTION

Velopharyngeal (VP) closure is important for normal speech production and swallowing. It enables establishment of normal resonance for non-nasal phoneme production and separation of the nasal cavity from the oropharynx to prevent nasal regurgitation during swallowing (Love & Webb, 2001). Dynamic movement and structural integrity of the soft palate (velum), lateral and posterior pharyngeal walls contribute to sufficient VP closure (Seikel, King & Drumright, 2000). Velopharyngeal inadequacy (VPI) is defined as disturbance to the velopharyngeal valving mechanism due to anatomical insufficiency or neurogenic incompetence, leading to inefficient segregation of nasal and oral cavities (Willging, 1999; Dworkin, Marunick, & Krouse, 2004; Johns, Rohrich & Awada, 2003).

Coupling of the nasal and oral cavities due to velopharyngeal dysfunctions may lead to the occurrence of articulation and resonance disorders (Peterson-Falzone, Hardin-Jones & Karnell, 2001). Hypernasality, nasal emission and weak pressure consonants are common speech characteristics in patients with VPI (Kummer, 2001; Shprintzen, & Bardach, 1995). Compensatory articulation might emerge in response to the difficulty in building up intra-oral pressure for production of pressure phonemes (Harding & Grunwell, 1998; Pullkkinen, Haapanen, Paaso, Laitinen, & Ranta, 2001). Velopharyngeal inadequacy commonly occurs in patients with cleft lip and palate and cleft palate (Hardin-Jones & Jones, 2005; Boseley & Hartnick, 2004; Inman, Thomas, Hodgkinson, & Reid, 2005). According to Daniller (1984), approximately 20% of children with repaired cleft lip and palate or cleft palate suffer from VPI. Conley, Gosain, Marks, & Larson (1997) reported that only 55% to 80% of the patients with primary palatal repair did not require secondary surgery for attainment of normal speech. Therefore, the impact of VPI in the cleft population should not be overlooked.

Speech therapy, prostheses and surgical management are advocated as the major treatment modalities for patients with VPI (Watson, Sell, & Grunwell, 2001; Marsh, 2003). Pharyngeal flap surgery and sphincter pharyngoplasty are the most common surgical procedures recommended to achieve VP closure (Ysunda, & Pamplona, 2005; Ysunza et al., 2004). However, positive outcome of secondary surgery is not guaranteed. Persistent hypernasality, hyponasality, difficulty in nasal breathing and, in the worst case, obstructive sleep apena, have been documented as the major postoperative complications (Potsic, Cotton & Handler, 1997; Marsh, 2003)

In order to achieve the best possible outcome for surgical management, a comprehensive assessment should be administered to gain an understanding of the possible causes and configuration of VP dysfunction. Assessment begins with perceptual speech evaluation and intra-oral examination, which gives preliminary information on the intactness and movement of the oral structures (Abdel-Haleem, 2003; Conley et al., 1997; Johns et al., 2003). Clinical examinations can only identify patients with possible needs for treatment, but is not sufficient to provide information for surgical management. The location, shape and size of VP gap, pattern and symmetry of VP closure should be considered for planning surgery (Daniller, 1984; Henningsson, & Isberg, 1991). Instrumental assessment complements the clinical examination by identifying the possible physical complications underlying the speech problems. Direct visualization of the VP mechanism during speech production is possible through instrumental assessment (Watson et al., 2001; Shprintzen, & Bardach, 1995). Therefore, it is an essential component for treatment planning.

Nasendoscopy is one of the most common imaging techniques used for the assessment of VPI (Rowe, & D'Antonio, 2005; D'Antonio, Achauer, & Vander Kam, 1993). Conley et. al. (1997) reported that nasendoscopy is utilized by up to 40% of

multidisciplinary cleft palate teams and the figure rises up to 90% for managing complex cases. Nasendoscopy allows direct visualization of the VP structures and their movements during speech production without physical interruption and exposure to radiation (Herrington, & Isberg, 1991). It provides information for deciding upon the management techniques and the possible needs for revision of the current condition of flaps or prostheses (D'Antonio, Muntz, Marsh, Marty-Grames, & Backensto-Marsh, 1988). In addition, the nasendscopy assessment could be videotaped for further analysis (Poppelreuter, Engelke, & Bruns, 2000). However, it is not without its limitations. Johns et al. (2003) mentioned that the invasiveness of nasendoscopy significantly affects its application, especially for young children. Havstam et al. (2005) reported that the nasendoscopy assessment was highly subjective and failed to give an accurate estimation of the size of VP gap or degree of closure.

The major means for reporting nasendoscopy assessment remains qualitative. Ramamurphy et. al (1997) mentioned that descriptive analysis is applied to estimate the relative contribution of the velopharyngeal valving mechanism and the symmetry of VP closure. Henningsson, & Isberg (1991) demonstrated the application of rating scales in describing the movement of velum, lateral and posterior pharyngeal walls in their study. Golding-Kushner et al. (1990) developed a standard protocol for reporting nasendoscopy and multiview videofluoroscopy assessment results. Estimation of the ratio of VP gap size and movements of lateral, posterior pharyngeal walls and velar during maximum closure during speech were suggested, with the scale ranging from 0.0 (no movement) to 1.0 (maximum movement). However, this standard method was still subjective.

Investigation of interrater and intrarater reliability is important to ensure the reliability and validity of assessment results for treatment planning. Pigott (2002) suggested that investigation of interrater and intrarater reliability is a must with a view to

the variability of nasendoscopy assessment interpretation across individuals. However, a limited amount of research was found in this aspect. D'Antonio, Marsh, Province, Muntz, & Phillips (1989) evaluated the reliability of perceptual rating of nasendoscopy images. Twelve raters were recruited to rate 125 video segments with a 6-point scale for estimating the relative movement of velar, lateral and posterior pharyngeal walls in their study. They were further divided into nine individual raters and an expert group of three raters. The results indicated that the reliability was higher for the expert group than individual raters. Yoon, Starr, Perkins, Bloom, & Sie (2007) investigated interrater and intrarater reliability of rating nasendoscopy images using the Golding-Kushner et al. (1990) scale. Six raters were recruited and asked to rate 50 nasendoscopy video segments for two times. They were required to estimate the ratio of gap size and velar, lateral and posterior pharyngeal wall movements with the scale ranging from 0.0 to 1.0. Satisfactory interrater and intrarater reliability were found for the use of this scale. Interrater and intrarater reliability in rating nasendoscopic assessment should be further investigated due to its importance for clinical decision making.

Several factors might interact with the reliability of rating nasendoscopy video segments. Fricatives and plosives are pressure phonemes, which are vulnerable to the effect of VPI (Shprintzen, & Bardach, 1995). It is unknown if the closure pattern or degree of closure differs in production of fricatives and plosives. If so, the reliability might be affected as the maximal closures or patterns of closure are different. No hypothesis is suggested here about the possible difference between the maximal closures or closure patterns in production of fricatives and plosives. However, the potential effect on reliability should be investigated as it may affect the decision making procedure for treatment planning. The choice of speech stimuli might also be an essential factor to consider.

The timing of assessment might affect the reliability. Nasendoscopy assessment is usually applied preoperatively and postoperatively to evaluate the VP mechanism (D'Antonio et al., 1988). The reliability in the application of the rating scale might differ in preoperative and postoperative assessments due to the change of velopharyngeal configuration. The VP gap is modified by artificial flaps or sphincters after the surgery. The gap size is significantly reduced at resting position and the relative degree of movement would be lower. It might impact on the variability in application of the scale among raters.

For the quality of nasendoscopy video signals, good quality images could not be guaranteed at any time. Some patients might be stimulated to produce more secretion during nasendoscopy assessment, masking the fiberoptic scope of nasendoscopy. In addition, it is sometimes difficult to visualize the VP structures due to anatomical constraints, for example, swelling of adenoid tissues (Witt, 1998). The scope might also become foggy during the application. Raters might have to make inference to the VP configuration and gap size. The variability of judgment might be higher among raters. The reliability of rating might be lower in condition of poor quality images.

In order to investigate the reliability of rating nasendoscopy assessments, VP gap size was chosen as the stimulus, rather than the velar, lateral or posterior wall movements. Firstly, the resonance of speech is directly determined by the VP gap size (Kummer, 2001). The velar, lateral and pharyngeal wall movements contribute to the VP closure, though. Secondly, the scope of investigation should be limited for thesis.

In conclusion, this study is designed to answer the following questions:

(1) What is the inter-rater reliability in rating velopharyngeal gap size? How well do the raters correlate with each other in application of the 6-point rating scale?

(2) What is the intra-rater reliability in rating velopharyngeal gap size?

(3) Do factors of speech samples used, timing and quality of video signals affect the reliability in rating velopharyngeal gap size using the 6-point scale?

i. How does the use of different speech stimuli (fricative and plosives) affect the interrater and intrarater reliability? Is there any difference in the reliability of rating velopharyngeal gap size in different conditions?

ii. How does the timing (preoperative and postoperative conditions) affect the interrater and intrarater reliability in rating VP gap size? As the resting position of postoperative condition is modified by either artificial flaps or sphincters, does it impact on the reliability between/within raters?

iii. How does the quality of nasendoscopy image impact on the interrater and intrarater reliability? Do the raters agree with each other better in condition with good quality videos? How does the reliability differ for these two conditions?

METHOD

Participants

Three expert raters participated in this study on a voluntary basis. Two of the raters were speech therapists while the other was an oral and maxillofacial surgeon. Two were Professors: one from the Division of Speech and Hearing Sciences and one from Oral and Maxillofacial Surgery of the University of Hong Kong. The third was a speech therapist with doctoral degree. Two had over ten years and one had less than eight years of experience in administration and interpretation of nasendoscopy examination.

Stimuli

Patients with repaired cleft lip and palate or cleft palate are followed up for speech evaluation in the Joint Cleft Lip and Palate Clinic in the Prince Philip Dental Hospital, the

University of Hong Kong. Patients with suspected velopharyngeal inadequacy are assessed by perceptual speech evaluation and instrumental assessment (flexible fiberoptic nasendoscopy and/or multiview videofluroscopy). Nasendoscopy is administered by the oral and maxillofacial surgeon, accompanied by the speech therapist. The flexible fiberoptic nasendoscopy is inserted through the nostril, and placed for visualization of velopharyngeal orifice. The Cantonese Nasendoscopy Speech Protocol (Whitehill, 2000) is used during the assessment which examines the integrity and movement of soft palate, lateral and posterior pharyngeal walls during speech production. Sentences loaded with plosives 'BB 俾波波爸爸' and '哥哥去街街買咯咯雞' and fricatives '四十一四十二四十三四十四四十五' with 16 and 15 syllables respectively, were chosen from the protocol for this study. The audio and video signals are recorded on VHS disk (from 1993 to 2000) or directly to computer (from 2000 till today). Written consent is obtained from patients or parents (for patients under 18 of age) for authorization in recording for research purpose.

A total of 19 subjects were included in this study. Their ages ranged from six to 27 (average: 14.2). All were patients with cleft lip and palate or cleft palate who had undergone either pharyngeal flap surgery or pharyngoplasty. The selection criteria were as follows: (1) Hearing 40 dB or better in at least one ear; (2) Preoperative and postoperative examinations should be undertaken within one year. The demographic details of the subjects are summarized in Table 1.

Table 1 Demographic details of the subjects

Subject	Sex	Age	Type of surgery
1	F	6	Sphincter pharyngoplasty
2	F	11	Sphincter pharyngoplasty
3	F	11	Sphincter pharyngoplasty
4	F	15	Sphincter pharyngoplasty
5	M	8	Sphincter pharyngoplasty
6	M	15	Sphincter pharyngoplasty
7	M	16	Sphincter pharyngoplasty
8	M	24	Sphincter pharyngoplasty
9	F	8	Pharyngeal flap surgery
10	F	11	Pharyngeal flap surgery
11	F	26	Phayngeal flap surgery
12	M	7	Phayngeal flap surgery
13	M	7	Phayngeal flap surgery
14	M	8	Phayngeal flap surgery
15	M	8	Phayngeal flap surgery
16	M	12	Pharyngeal flap surgery
17	M	24	Pharyngeal flap surgery
18	M	27	Pharyngeal flap surgery
19	F	25	Lateral pharyngoplasty

Preoperative and postoperative examination video recordings were selected for each subject. The VHS recording was converted to DVD format for editing. Each video recording was edited by Ulead VideoStudio 9.0 SE DVD®. Video recordings during the production of sentences loaded with bilabial plosives ‘BB 俾波波爸爸’ and velar plosives ‘哥哥去街街買咯咯雞’ and fricatives ‘四十一四十二四十三四十四四十五’ were selected. The number of syllables in sentences loaded with plosives and fricatives were similar. Four video clips were prepared for each subject:

(1) Sentences loaded with bilabial and velar plosives (Preoperative recording)

(2) Sentence loaded with fricatives (Preoperative recording)

(3) Sentences loaded with bilabial and velar plosives (Postoperative recording)

(4) Sentence loaded with fricatives (Postoperative recording)

For the factor of quality of samples, nasendoscopy video segments were classified into two categories: good quality samples and poor quality samples by the author. The criteria for good quality sample were set as follows: (1) The image should not be masked by fog; (2) Each anatomical structure of the VP sphincter should be clearly shown.

Procedures

A 6-point rating scale was used for the rating of VP gap size. The scale was adapted from Golding-Kushner et al. (1990). Golding-Kushner et al. (1990) suggested that VP gap size should be reported by the ratio of maximum closure relative to the resting position. The raters were required to provide a rating between 0.0 to 1.0. No closure movement was rated as 0.0 and maximal closure was rated as 1.0. For the 6-point scale employed in this study, only six points were provided for rating and descriptions were provided for three scale points only in order to minimize bias. The 6-point scale differed from the reporting method suggested by Golding-Kushner et al. (1990) in the way that the number of points was restricted for rating VP gap size. The scale was developed and employed in the study by Chanchareonsook, Whitehill, & Samman (2007). There were two reasons for using the scale in the current study. Firstly, the reliability of rating VP gap size using the scale suggested by Golding-Kushner et al. (1990) had been examined previously. Yoon et al. (2006) investigated the interrater and intrarater reliability using the scale of Golding-Kushner et al. (1990). Secondly, the variability of the 6-point scale was lower than that of the scale suggested by Golding-Kushner et al.(1990). The 6-point scale had been used

previously in the study by Chanchareonsook, Whitehill, & Samman (2007). However, its reliability had not been investigated yet.

Table 2 6-point rating scale (based on Golding-Kushner et. al., 1990)

Rating	Description
0.0	No movement
0.25	
0.50	
0.75	
0.90	Borderline/pin hole closure
1.0	Complete VP closure

A computer programme was developed and run by E-prime for the rating session. Each video clip was rated twice in order to measure the intra-rater reliability. The video clips were randomized in order of presentation.

A briefing session was provided to familiarize the raters with the background and procedures of the experiment. The raters were advised to attempt to give their best rating regardless of the quality of the audio and video signals. For subjects with more than one VP port (after pharyngeal flap surgery), the rater was informed to use the port with better closure for rating. A screening block followed the briefing session. The raters were guided to give their rating for four trials to ensure smooth administration.

The rating session was divided into two parts. In each part, there were four blocks; each block consisted of 19 trials. The raters were free to take a break between each block and part. The whole rating session took about 2 hours. A face-to-face interview was

administered after the rating session to investigate if there were any factors which might affect the raters' performance.

Reliability and agreement measures were applied to answer the research questions. Interrater and intrarater reliability were calculated by intraclass correlation coefficient (ICC) and Pearson correlation coefficient, respectively (Kreiman & Gerratt, 1993). Exact agreement and agreement within one scale point were calculated manually.

For exact agreement, the ratings of the raters should be the same for a single trial; for agreement within 1 scale point, the ratings should not differ from each other more than one scale value. For instance, if two of the raters gave a rating of 0.9 and one gave a rating of 1.0 in the first trial, this trial does not meet the criteria for exact agreement, but fit the requirement of agreement within one scale point. If the raters gave ratings of 0.5, 0.75 and 0.9 for a single trial, this trial meets neither the requirements for exact agreement nor agreement within one scale point. The criterion for exact agreement was strict. Interpretation of the research results using exact agreement should be careful.

RESULTS

Interrater reliability and agreement

Intraclass correlation coefficient was calculated in order to investigate how well the raters correlated with each other in estimating velopharyngeal gap size. The interrater reliability was calculated using the 76 trials given by each rater. The interrater reliability was 0.763 ($p < 0.05$).

Interrater agreement was computed using exact agreement and agreement within one scale point. The agreement within one scale point was 75% and exact agreement was 26%.

Intrarater reliability and agreement

The degree of consistency within each rater was also calculated. Each rater estimated the gap size for each stimulus twice. Pearson's correlation coefficient was applied to calculate the intrarater reliability for each rater. The intrarater reliability ranged from 0.421 to 0.739 for the three raters.

Agreement values were also calculated to investigate if each rater applied and interpreted the rating scale consistently. The agreement within one scale point and exact agreement were computed. The agreement within one scale point and exact agreement varied from 76% to 97% and 46% to 55%, respectively.

Table 3 Intrarater reliability and agreement in estimation of VP gap size

	Intrarater reliability coefficient	**% of agreement**	
		Exact	Within 1 scale point
Rater 1	.739 ($p < 0.05$)	55% (42/77)	97% (75/76)
Rater 2	.735 ($p < 0.05$)	55% (42/76)	89% (68/76)
Rater 3	.421 ($p < 0.05$)	46% (35/76)	76% (58/76)

Speech samples used

In an attempt to investigate if the reliability of rating is affected by different speech stimuli, interrater and intrarater agreement and reliability were calculated. Interrater reliability was 0.836 ($p < 0.05$) for speech samples loaded with plosives and 0.676 ($p < 0.05$) for speech samples loaded with fricatives. For interrater agreement, agreements within one scale point were 82% and 68% for speech samples of plosives and fricatives, respectively while the exact agreement was 26% in both conditions.

Intrarater agreement and reliability in conditions with sentences loaded with plosives and fricatives were also calculated. For intrarater reliability, the reliability values ranged from 0.389 to 0.796 for plosives and 0.452 to 0.722 for fricatives. For intrarater agreement, the agreement within one scale point ranged from 71% to 97% for plosives and 82% to 97% for fricatives; the exact agreement varied from 34% to 58% for plosives and 53% to 58% for fricatives. The differences in intrarater agreement and reliability values were higher for plosives.

Table 4 Intrarater agreement and reliability in estimation of VP gap size in conditions of sentences loaded with plosives and fricative

	Type of agreement measure	Rater 1	Rater 2	Rater 3
Plosives	Reliability	0.774 ($p < 0.05$)	0.796 ($p < 0.05$)	0.389 ($p < 0.05$)
	% of exact agreement	55%	58%	34%
	% of agreement within 1 scale	97%	92%	71%
Fricatives	Reliability	0.722 ($p < 0.05$)	0.671 ($p < 0.05$)	0.452 ($p < 0.05$)
	% of exact agreement	55%	53%	58%
	% of agreement within 1 scale	97%	87%	82%

Timing

The second factor investigated was the potential effect of timing. The ratings of preoperative and postoperative assessments were compared by computation of interrater and intrarater agreement and reliability.

The interrater reliability values for preoperative and postoperative assessments were 0.721 ($p < 0.05$) and 0.618 ($p < 0.05$) respectively. The interrater agreements within one scale point were 74% for preoperative condition and 76% for postoperative condition; the exact agreements were 26 % in both conditions. The interrater agreement values, including the exact agreement and agreement within one scale point, were similar across preoperative and postoperative conditions.

Intrarater agreement and reliability values were also calculated. For intrarater reliability, the reliability values ranged from 0.546 to 0.682 in preoperative condition and from 0.172 to 0.709 in postoperative condition. For agreement values, the agreement within one scale point varied from 82% to 100% in preoperative condition and from 68% to 95% in postoperative condition; the exact agreement ranged from 47% to 55% in preoperative condition and 45% to 63% in postoperative condition. The difference in intrarater agreement and reliability values among raters was higher in postoperative condition. The intrarater reliability was insignificant for one of the raters in postoperative assessment.

Table 5 Intrarater agreement and reliability in estimation of VP gap size in preoperative and postoperative assessments

	Type of agreement measure	Rater 1	Rater 2	Rater 3
Preoperative	Reliability	0.682 ($p < 0.05$)	0.546 ($p < 0.05$)	0.607 ($p < 0.05$)
	% of exact agreement	47%	55%	47%
	% of agreement within 1 scale	100%	87%	82%
Postoperative	Reliability	0.610 ($p < 0.05$)	0.709 ($p < 0.05$)	0.172 ($p > 0.05$)
	% of exact agreement	63%	55%	45%
	% of agreement within 1 scale	95%	92%	68%

Quality of nasendoscopic images

In order to evaluate the effect of quality of nasendoscopy images on reliability, interrater and intrarater agreement and reliability were measured in conditions of good quality images and poor quality images.

For interater reliability, the reliability values were similar across the two conditions. The reliability values were 0.797 ($p < 0.05$) for good quality image and 0.779 ($p < 0.05$) for poor quality image. For the agreement values, the agreement within one scale point was 71% and 79% for good and poor quality images; while the exact agreements were 29% for good quality images and 21% for poor quality images.

For intrarater reliability, the reliability values ranged from 0.621 to 0.780 and 0.153 to 0.797 in conditions with good and poor quality images, respectively. For intrarater agreement, the agreements within one scale point varied from 82% to 96% in condition with good quality images and 71% to 96% with poor quality images. The exact

agreements ranged from 50% to 57% for good quality images and 46% to 61% for poor quality images. The difference in intrarater agreement and reliability values was larger in condition with poor quality images. The intrarater reliability was insignificant for one of the raters in condition with poor quality images.

Table 6 Intrarater agreement & reliability in estimation of VP gap size in conditions with good and poor quality images.

	Type of agreement measure	Rater 1	Rater 2	Rater 3
Good quality	Reliability	0.731 ($p < 0.05$)	0.621 ($p < 0.05$)	0.780 ($p < 0.05$)
	% of exact agreement	57%	50%	54%
	% of agreement within 1 scale	96%	82%	89%
Bad quality	Reliability	0.686 ($p < 0.05$)	0.797 ($p < 0.05$)	0.153 ($p > 0.05$)
	% of exact agreement	46%	61%	54%
	% of agreement within 1 scale	96%	93%	71%

DISCUSSION

Interrater reliability

The first purpose of this study was to investigate how well the raters correlated with each other in rating velopharygneal gap size. Interrater reliability was higher (ICC= 0.763) than that reported by Yoon, e.t.al. (2006). In addition, the interrater agreement within one scale point was also high.

In the study by Yoon et. al.(2006), the interrater reliability was 0.57 among six raters, including two faculty otolaryngologists, two pediatric otolaryngology fellows and two speech pathologists, for estimation of the VP gap size using the Golding-Kushner et al. (1990) scale. They also determined the interrater reliability between the two faculty otolaryngologists, the two pediatric otolaryngology fellows and the two speech pathologists. The reliability values ranged from 0.43 to 0.63, which were much lower than the results of this study.

The discrepancy of interrater reliability among raters in this study compared with Yoon, et. al.(2006) might be related to the difference in the rating scale. Yoon, et.al. (2006) adapted the Golding-Kushner et al. (1990) scale directly for estimation of velophayrngeal gap size. The percentage of the velopharyngeal closure relative to the resting position was used and this scale ranged from 0 to 100%. There was a lot of freedom for interpretation between the two end points of this scale. But for the 6-point scale used in our current study, the interpretation of the scale points was relatively limited. Therefore, the reliability between the raters was higher in this current study.

In this study, the three raters correlated well with each other in the application of the 6-point scale. It revealed that their interpretation and understanding of the points of scale were fairly similar. For communication of assessment findings in the management team, the personnel involved, including the speech therapist and dental surgeon, should be able to understand well the degree of severity for planning appropriate procedures and the size of flaps or sphincters. Therefore, this finding suggested that the use of the six point scale was reliable among the three raters. As the raters worked together in the Joint Cleft Lip and Palate Clinic for nasendoscopy assessment, this finding provided evidence of the reliability of the nasendoscopy assessment in this clinic.

Intrarater reliability

The second purpose of this study was to investigate the consistency of each rater in estimation of velopharyngeal gap size using the 6-point scale. The intrarater reliability ranged from 0.42 to 0.74. The agreement (within one scale point) of the raters ranged from 76% to 97%. In the study by Yoon et al. (2006), the intrarater reliability varied from 0.66 to 0.94 for all 6 six raters. Agreement measure was not applied in the study of Yoon et al. (2006). Compared with this study, the intrarater reliability of raters in Yoon et al.'s study (2006) was higher.

Experience of application might be a potential factor for explaining the discrepancy of intrarater agreement and reliability values between raters in current study. Rater 3 had the lowest agreement (within 1 scale point) and reliability values among the raters. It implied that the consistency in the application of the scale was much lower than other raters. It was reported that the raters had over 10 years of experience in working with nasendoscopy assessment, except rater 3. Rater 3 had less than eight years of experience in using nasendoscopy for assessment. In addition, the familiarity with the 6-point scale might be another factor. The 6-point rating scale was developed by rater 1 and 2 and used in the study by Chanchareonsook, Whitehill, & Samman (2007). They should have better understanding of the meaning of each scale point and acquire higher consistency in rating due to their knowledge about the scale. Yoon et al. (2006) also mentioned the potential effect of experience on the reliability of rating nasendoscopy assessment. In their study, the raters who had the least experience with nasendoscopy assessment exhibited the lowest reliability values.

Speech samples

The third purpose of this study was to determine if any factors might affect the agreement and reliability of raters in application of the 6-point scale. The interrater agreement and reliability values for rating speech samples loaded with plosives and fricatives were compared. The reliability values were 0.84 for plosives and 0.68 for fricatives. The agreement within one scale point was 82% for plosives and 68% for fricatives. The agreement and reliability values for plosives were higher than those of fricatives.

The possible explanation for the lower agreement for segments with fricatives might be related to the anatomical and physiological difference between the production of fricatives and plosives. Plosives are produced with sudden episodes of closure and opening of the velopharyngeal sphincter. For fricatives, continuous and sustained closure pattern of velopharyngeal sphincter are required for the direction of airflow to the oral cavity to create the frication noise. The raters were instructed to give the ratio of the maximum closure relative to the resting position. The maximum closure might be more difficult to be detected for fricative production due to continuous closure of velopharyngeal sphincter. In contrast, the best closure for plosives might be more easily identified due to its sudden nature. The consistency might be affected due to the difficulty in identifying the maximum velopharyngeal closure for fricatives for different raters. No previous research explored the effect of speech samples used on the reliability of rating.

Intrarater agreement and reliability values were also computed. But the results did not support the hypothesis suggested above. The intrarater agreement values (within one scale point) for rater 1 and 2 were high and similar for fricatives and plosives. For rater 3, the intrarater agreement value (within one scale point) for fricatives was higher than that of plosives. The intrarater reliability values were high and similar for fricatives and

plosives for rater 1. However, the reliability value for plosives was higher than that of fricatives for rater 2. On the contrary, the reliability value for plosives was lower than that of fricatives for rater 3, but the correlation values are relatively insignificant in both conditions. The results were inconsistent and no conclusion could be made about the possible effect of speech samples on the agreement and reliability of rating.

Timing

The interrater and intrarater agreement and reliability were calculated to compare if there was any difference between preoperative and postoperative assessments.

Interrater reliability was 0.72 for preoperative condition and 0.62 for postoperative condition. No difference was observed for agreements within one scale point and exact agreements between the preoperative and postoperative conditions. For intrarater agreement (within one scale point), no remarkable difference was observed for rater 1 and 2 in different conditions and the agreement values (within one scale point) were high. No distinct difference was observed between the reliability values of preoperative and postoperative assessments for rater 1. Rater 2 demonstrated lower reliability value for preoperative assessment. Rater 3 exhibited satisfactory agreement (within one scale point) and fair reliability value for preoperative assessment. The intrarater reliability value for rater 3 was insignificant and the agreement within one scale point was fair in condition of postoperative assessment. The discrepancy between preoperative and postoperative assessments was more explicit for reliability than agreement value (within one scale point) for rater 3.

The interrater reliability was higher for preoperative assessments in application of the scales than that of the postoperative assessments. It implied that there was lower consistency in application of the 6-point scale in rating VP gap size between raters in

condition of postoperative assessment. In postoperative assessments, flaps or sphincters reduced the gap size. As the configuration of the resting position of the VP gap was completely different and significantly reduced in size, the rating of the ratio of maximum VP closure relative to the resting position might be more difficult. The interpretation of the scale points might differ in greater extent between raters.

For intrarater reliability, the value of postoperative assessment was insignificant for rater 3. It might be due to the difficulty to achieve consistent application of rating scales for estimating VP gap size in postoperative assessment. Experience with the use of nasendoscopy might be another factor which might affect the consistency in application of rating scales. Rater 1 and 2 gave comparably acceptable agreement and reliability in rating velopharyngeal gap size in postoperative condition. They both had over 10 years of experience in application of the scales and were involved in the development of this scale. Therefore, their application was more consistent, even in unfavourable condition.

However, rater 2 yielded better reliability in postoperative assessment than preoperative assessment. This result was quite contradictory to that of the hypothesis above. Therefore, it is concluded that the velophaynrgeal configuration in preoperative and postoperative conditions might not be a possible factors which would affect the reliability.

Quality of nasendoscopic images

The interrater and intrarater agreement and reliability values were calculated to determine if the quality of the video clips posed significant effect on the reliability of the raters.

The interrater reliability values, agreement within one scale, and exact agreement were comparable for variables of good quality and bad quality nasendoscopic video clips.

For intrarater reliability, contradictory results were identified. Rater 2 demonstrated higher agreement (within one scale point) and reliability for rating poorer video clips; Rater 3 exhibited significantly lower reliability value for poor quality videos than good quality videos. The application of rating scale was highly inconsistent in poor quality condition within rater 3.

All three raters reported that the quality of the audio signals of the video clips affected their rating. The articulation errors and the speech quality might cause bias on the rating. The raters were instructed to attempt to give the rating regardless of the quality of the nasendoscopy video segments. However, some raters might be more vulnerable to the effect of the speech quality. The effect of the speech quality on reliability might be relatively individualized. Further research is needed to make conclusion about the possible effect of this factor. It is highly related to the clinical application of nasendoscopy. In nasendoscopy assessment, the patients are required to produce speech samples to examine the velopharyngeal movement. If the raters' judgment is biased due to the speech quality, the assessment results might not be accurate and reliable for decision making for clinical purpose.

In conclusion, good quality video is preferred for rater with relatively less experience in nasendoscopy assessment. In this study, highly consistent application of rating scales was revealed for experienced raters. Raters with less experience were prone to the effect of external factors.

General discussion

For clinical purpose, agreement among the multidisciplinary management team for cleft lip and palate is critical. They should have similar understandings of the definition of each scale point of the rating system, which encourages proper communication of

assessment findings between the personnel for accurate decision making, such as the location or size of flaps or sphincters.

For research purpose, the reliability between the raters should be carefully considered. The raters' reliability in application of the rating scale in regular fashion enables the consistency of findings for studying some phenomena, for example, the relationship between manner of articulation and VP closure pattern or degree of VP closure.

As the quality of the nasendoscopy video cannot always be ensured, the experience of the raters become an important factor for accurate ratings of velophayngeal gap size for both clinical and research purpose. In order to enhance the raters' competence in rating VP closure at any condition even if the quality of the nasendoscopy segments are far from satisfactory, training session might be a possible suggestion for increasing the exposure of raters to nasendoscopy images upon feedback provision. Training session was advocated by Yoon, et al.(2006) to enhance the interrater reliability in rating velopharyngeal movement.

Limitations of the study

Due to the limited number of experts raters recruited, the external validity of this study could not be ensured.

Conclusion and clinical implications

In conclusion, there was satisfactory interrater and intrarater agreement and reliability in using a six point rating scale to estimate velopharyngeal gap size. Nasendoscopy is a reliable tool for assessing velopharyngeal function for clinical and research purposes.

ACKNOWLEDGEMENTS

I would like to express my thankfulness to Professor Tara Whitehill for her sincere guidance and support for the preparation of this dissertation project. I would like to thank Dr Bradley McPherson for his support during the period of leave of Professor Whitehill, Professor Samman for his kind participation in my study through the provision of nasendoscopy stimuli and being one of the raters, Dr Karen Chan for her assistance in preparation of the tailor-made computer program for the rating session, Dr Joyce Chun for her keen participation as the rater. Last but not least, I would like to thank Mr Raymond Wu and Mr Donald Chan for providing technical support for the procedure of data collection.

REFERENCES

Abdel-Haleem, E.K. (2003). Protocol of assessment of velopharyngeal incompetence. *International Congress Series, 1240,* 663-667.

Boseley, M.E., & Hartnick, C.J. (2004). Assessing the outcome of surgery to correct velopharyngeal insufficiency with the pediatric voice outcomes survey. *International Journal of Pediatric Otorhinolaryngology, 68,* 1429-1433.

Chanchareonsook, N., Whitehill, T.L., & Samman, N. (2007). Speech Outcome and Velopharyngeal Function in Cleft Palate: Comparison of Le Fort I Maxillary Osteotomy and Distraction Osteogenesis-Early Results. *The Cleft Palate - Craniofacial Journal, 44(1),* 23-32.

Conley, S.F., Gosain, A.K., Marks, S.M., & Larson, D.L. (1997). Identification and assessment of velopharyngeal inadequacy. *American Journal of Otolaryngology, 18(1),* 38-46.

D'Antonio, L.L., Achauer, B.M., & Vander Kam, V.M. (1993). Results of a survey of cleft palate teams concerning the use of nasendoscopy. *Cleft Palate-Craniofacial Journal, 30(1),* 35-39.

D'Antonio, L.L., Marsh, J.L., Province, M.A., Muntz, H.R., & Philips, C.J. (1989). Reliability of flexible fiberoptic nasopharyngoscopy for evaluation of velopharyngeal function in a clinical population. *Cleft Palate Journal, 26(3),* 217-225.

D'Antonio, L.L., Muntz, H.R., Marsh, J.L., Marty-Grames, L., & Backensto-Marsh, R. (1988). Practical application of flexible fiberoptic nasopharyngoscopy for evaluating velopharyngeal function. *Plastic and Reconstructive Surgery, 611-618.*

Daniller, A. (1984). Use of nasendoscopy in the treatment of velopharyngeal insufficiency. *The Western Journal of Medicine, 141(2),* 232-233.

Dworkin, J.P., Marunick, M.T., & Krouse, J.H. (2004). Velopharyngeal dysfunction: Speech characteristics, variables etiologies, evaluation techniques, and differential treatment. *Language, Speech & Hearing Services in Schools, 35(4),* 333-352.

Golding-Kushner, K.J., Argamaso, R., Cotton, R., Grames, L., Henningsson, G., Jones, D., Karnell, M., Klaiman, P., Lewin, M., Marsh, J., McCall, G., McGrath, C., Muntz, H., Nevdahl, M., Rakoff, S., Shprintzen, R., Sidoti, E., Vallino, L., Volk, M., Williams, W., Witzel, M.A., Dixon Wood, V.L., Ysunza, A., D'Antonio, L., Isberg, A., Pigott, R., & Skolnick, L.(1990). Standardization for the reporting of nasopharyngoscopy and multiview videofluoroscopy: A report from an international working group. *Cleft Palate Journal, 27,* 337-348.

Harding, A., & Grunwell, P. (1998). Active vs passive cleft-type speech characteristics. *International Journal of Language and Communication Disorders, 33,* 329-352.

Hardin-Jones, M.A., & Jones, D.L. (2005). Speech production of preschoolers with cleft palate. *Cleft Palate-Craniofacial Journal, 42(1),* 7-13.

Havstam, C., **Lohmander, A., Persson, C., Dotevall, H., Lith, A., & Lilja, J.** (2005). Evaluation of VPI-assessment with videofluoroscopy and nasoendoscopy. *British Jorunal of Plastic Surgery, 58(7),* 922-931.

Henningsson, G., & Isberg, A. (1991). Comparison between multiview videofluoroscopy and nasendoscopy of velopharyngeal movements. *Cleft Palate-Craniofacial Journal, 28(4),* 413-418.

Inman, D.S., Thomas, P., Hodgkinson, P.D., & Reid, C.A. (2005). Oro-nasal fistula development and velopharyngeal insufficiency following primary cleft palate surgery - an audit of 148 children born between 1985 and 1997. *British Journal of Plastic Surgery, 58,* 1051-1054.

Johns, D.F., Rohrich, R.J., & Awada, M. (2003). Velopharyngeal incompetence: A guide for clinical evaluation. *Journal of the American Society of Plastic and Reconstructive Surgery, 112(7),* 1890-1898.

Kreiman, J., & Gerratt, B.R. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech & Hearing Research, 36(1),* 21-57.

Kummer, A.W. (2001). *Cleft palate and craniofacial anomalies: The effects on speech and resonance.* San Diego: Singular Thomson Learning.

Love, R.J., & Webb, W.G.(2001). *Neurology for the speech-language pathologist (4th ed.).* Boston: Butterworth-Heinemann.

Marsh, J.L. (2003). Management of velopharyngeal dysfunction: Differential diagnosis for differential management. *The Journal of Craniofacial Surgery, 14(5),* 621-628.

Peterson-Falzone, S.J., Hardin-Jones, M.A., & Karnell, M.P. (2001). *Cleft palate speech.(3rd ed.).* St Louis: Mosby.

Pigott, R.W. (2002). An analysis of the strengths and weaknesses of endoscopic and radiological investigation of velopharyngeal incompetence based on a 20 year experience of simultaneous recording. *British Journal of Plastic Surgery, 55,* 32-34.

Poppelreuter, S., Engelke, W., & Bruns, T. (2000). Quantitative analysis of the velopharyngeal sphincter function during speech. *Cleft Palate-Craniofacial Journal, 37(2),* 157-165.

Potsic, W.P., Cotton, R.T., & Handler, S.D. (1997). *Surgical pediatric otolaryngology.* New York: Thieme.

Pulkkinen, J., Haapanen, M-L., Paaso, M., Laitnen, J., & Ranta, R. (2001). Velopharyngeal function from the age of three to eight years in cleft palate patients. *Folia Phoniatrica et Logopaedia,* 53, 93-98.

Ramamurphy, L., **Wyatt, R.A., Whitby, D, Martin, D., & Davenport, P.** (1997). The evaluation of velopharyngeal function using flexible nasendoscopy. *The Journal of Laryngology and Otology, 111,* 739-745.

Rowe, M.R., & D'Antonio, L.L. (2005). Velopharyngeal dysfunction: Evolving developments in evaluation. *Lippincott Williams & Wilkins, 13,* 366-370.

Seikel, J.A., King, D.W., & Drumright, D.G. (2000). *Anatomy and physiology for speech, language, and hearing. (2nd ed.).* San Diego: Singular Publishing Group.

Shprintzen, R.J., & Bardach, J. (1995). *Cleft palate speech management: A multidisciplinary approach.* St. Louis: Mosby.

Watson, A.C.H., Sell, D.A., & Grunwell, P. (2001). *Management of cleft lip and palate.* London: Whurr Publishers.

Whitehill, T.L. (2000). *Cantonese Nasendoscopy Speech Protocol.*

Willging, J.P. (1999). Velopharyngeal insufficiency. *International Journal of Pediatric Otorhinolaryngology*, S307-S309.

Witt, P.D. (1998). Evaluation of the velopharynx: Past, present, future. *European Journal of Plastic Surgery, 21(3),* 123-128.

Yoon, P.J., Starr, J.R., Perkins, J.A., Bloom, D., & Sie, K.C.Y. (2006). Interrater and intrarater reliability in the evaluation of velopharyngeal insufficiency within a single institution. *Archives of Otolaryngology - Head and Neck Surgery, 132*, 947-951.

Ysunza, A., & Pamplona, C. (2005). Velophayrngeal function after two different types of pharyngoplasty. *International Journal of Pediatric Otorhinolaryngology, 70,* 1031-1037.

Ysunza, A., Pamplona, C., Fernando, M., Drucker, M., Felemovicius, J., Ramirez, E., & Patino, C. (2004). Surgery for speech in cleft palate patients. *International Journal of Pediatric Otorhinolaryngology, 68*, 1499-1505.