Interrater and Intrarater Reliability in the Evaluation of Velopharyngeal Insufficiency Within a Single Institution

Patricia J. Yoon, MD; Jacqueline R. Starr, PhD; Jonathan A. Perkins, DO; David Bloom, MD; Kathleen C. Y. Sie, MD

Objective: To explore the interrater and intrarater reliability in nasoendoscopic assessment of velopharyngeal (VP) function using the standardized reporting method described by Golding-Kushner within a single institution.

Design: Prospective blinded study.

Setting: Academic, tertiary care, pediatric hospital.

Participants: Six health care providers (2 pediatric otolaryngology faculty members, 2 pediatric otolaryngology fellows, and 2 speech pathologists) independently rated 50 videotaped nasoendoscopy segments twice. The segments on the videotape were obtained in a clinical setting.

Main Outcome Measures: The Golding-Kushner rating system was used to rate VP function. Raters described VP closure quantitatively by rating palatal and lateral pharyngeal wall movement for each segment. They also qualitatively described characteristics of the VP gap, rated gap size as none, small, medium, or large, and estimated the percentage gap size relative to the resting position. Reliability coefficients were calculated for the data sets.

Results: Fairly good interrater and intrarater reliability was seen in the quantitative measures. Faculty otolaryngologists rated segments more similarly to each other than did pediatric otolaryngology fellows, but intrarater reliability was similar for both the experienced and less experienced otolaryngologists. Less consistency was seen in the ratings of the speech pathologists. Raters tended to rate with less consistency when describing qualitative characteristics of the VP gap than when making quantitative measurements.

Conclusions: The Golding-Kushner scale is a reasonably reliable tool for reporting nasoendoscopic findings at our institution. However, these data also indicate that there exists room for improvement and that rater training may increase reliability.


VELOPHARYNGEAL (VP) CLOSURE is required for normal speech production. Incomplete closure of the velopharynx during speech results in VP insufficiency (VPI), characterized by hypernasal resonance and nasal air escape. Velopharyngeal insufficiency may compromise speech intelligibility and may play a causal role in hoarseness and development of compensatory misarticulations.

Velopharyngeal insufficiency is most commonly associated with cleft palate. However, it may also be associated with other causes, including neuromuscular conditions and idiopathic causes. Regardless of etiologic origin, degree of VP closure can vary considerably between patients. Surgical treatment of VPI poses the challenge of eliminating nasal air escape during speech while preserving an adequate nasal airway during respiration. Many authors have acknowledged the importance of "tailoring" the intervention to the patient's needs.1-4 Treatment of VPI requires an understanding of its causes and careful assessment of the degree and shape of the VP closure.

Assessment of VP function starts with careful perceptual speech analysis because it is the degree of speech compromise that dictates the need for intervention. Instrumental assessment then provides information regarding VP gap size and shape that can be helpful in making surgical recommendations. Commonly used methods for assessment include nasoendoscopy and multiview videofluoroscopy. Both examinations are designed to assess the degree of palatal and lateral pharyngeal wall movement during speech. A
number of reporting schemes have been proposed in the past. In 1990, a multidisciplinary international working group proposed a standardized rating system for the reporting of instrumental assessments of VPI. Using this scale, which we will call the Golding-Kushner scale, the maximum movement of the lateral pharyngeal walls and the muscular palate during speech are rated relative to the resting position. This scale can be applied to both nasoendoscopy and videofluoroscopy.

The application of this rating scale is subjective, and it is therefore important to understand the interrater and intrarater variability. As we attempt to further refine our treatment approaches to optimize speech outcomes and minimize complications, such as persistent VPI or airway obstruction, it is important for us to understand the degree of variability in the measurement tools. This understanding is particularly critical because there is no gold standard objective measure of degree of VP closure during speech.

Multiple authors have investigated the relationship of the endoscopic and fluoroscopic assessments of VP function. In addition to understanding the relationship between the different types of instrumental assessments of VP function, it is also important for health care providers to understand the variability of ratings within each type of diagnostic evaluation. Minimizing variability will facilitate the standardization of reporting of the degree of VP dysfunction. Standardization, in turn, will help us to refine treatment protocols within a single institution and is critical in designing multicenter protocols.

The objective of this study is to assess the interrater and intrarater reliability of the standardized Golding-Kushner reporting system as applied to flexible fiberoptic nasoendoscopic assessment of VP function at our institution. We compared the ratings of 6 raters, including 2 attending pediatric otolaryngologists, 2 pediatric otolaryngology fellows, and 2 pediatric speech pathologists.

### METHODS

The interrater and intrarater reliability study was designed to assess measurement error in rating previously videotaped segments of nasoendoscopic examinations. Institutional review board exemption (E03-115-01) was granted on June 5, 2003.

### PATIENTS AND PROCEDURES

Patients referred to Children's Hospital and Regional Medical Center for speech delays are evaluated by a pediatric speech pathologist. Patients who have VPI on perceptual speech analysis are referred for instrumental assessment of VP function and routinely undergo nasoendoscopy. Endoscopic examinations are performed by an otolaryngologist with a speech pathologist in attendance. A flexible fiberoptic nasoendoscope is introduced through the middle meatus, and patients are prompted to give a speech sample. These examinations, including audio input using a directional microphone, were recorded onto Super VHS tapes, and a library of such tapes was maintained. The study segments were selected from among tapes of all patients with VPI evaluated in this clinic. Only segments containing a representative speech sample were included. The segments were deidentified and labeled from 1 to 30 on a single tape.

Four otolaryngologists and 2 speech pathologists each rated the endoscopic segments twice. Raters were instructed to view the videotape and rate the lateral pharyngeal wall and palatal movement based on the Golding-Kushner scale. Lateral pharyngeal wall movements were rated on a scale from 0 to 0.5 on each side, with 0 indicating absence of medial excursion of the lateral pharyngeal wall and a score of 0.5 indicating movement to the midline. Palatal elevation was rated from 0 to 1.0 on the right and left sides, with 0 indicating absence of palatal elevation on phonation and 1.0 indicating normal palatal contact with the posterior pharyngeal wall. The posterior pharyngeal wall was also rated from 0 to 1.0, with 0 indicating absence of posterior pharyngeal wall movement and 1.0 indicating anterior movement to meet the velum. Each segment was rated for presence or absence of aberrant pharyngeal wall pulsations, the Passavant ridge, and palatal notching. Finally, raters were asked to assign a gap size (none, small, medium, or large) and percentage of gap size relative to resting gap size. A rating form was provided (Figure).

Raters were asked to review and rate the tapes individually without consulting each other. Each segment could be viewed as many times as necessary, but the raters were instructed not to rewind and rerate prior segments. Raters were blinded from each others' ratings. To evaluate intrarater reliability, a second rating of the tape was performed several weeks later in an identical manner to the first. The raters were blinded from their previous ratings.

### STATISTICAL ANALYSIS

Reliability of the ratings was assessed for the following components of VP movement: (1) right lateral wall movement plus left lateral wall movement; (2) ½ right palatal movement plus ½ left palatal movement; (3) estimated percentage of gap size relative to the resting position.

Interrater and intrarater reliability coefficients were calculated from 2-way random effects analysis of variance models. The model allowed for the simultaneous assessment of the intrarater reliability coefficient (averaged over the first and second ratings for each rater) and intrarater reliability coefficient (averaged over all raters), while accounting for the correlation induced by the repeated measures. The reliability co-
efficient can range from 0 to 1.0 and indicates the proportion of variability of a continuous measurement attributable to true differences across segments. When the reliability coefficient is high, interrater and intrarater variability are relatively low compared with other sources of variability, such as actual differences in gap size.

The $\kappa$ coefficient was calculated to determine variability in evaluating the presence of the Passavant ridge, palatal midline notch, and aberrant pulsations, and in the assignment of gap size category (none, small, medium, or large). The $\kappa$ coefficient measures agreement between 2 independent categorical ratings beyond that expected due to chance. Similar to the reliability coefficient, the $\kappa$ coefficient may take on values from 0 to 1.0, where 0 indicates that any agreement observed is completely by chance and 1.0 indicates complete agreement between observers.

All analyses were performed using STATA version 8.0 (Stata-Corp LP, College Station, Tex).

### RESULTS

The 50 video segment lengths ranged from 26 seconds to 4 minutes 36 seconds, with a median length of 80 seconds. The length of the videotaped segments was not associated with either the interrater or intrarater reliability.

The videotape included evaluations of 20 females and 30 males, ranging in age from 3.5 to 17 years, with a mean±SD age of 9.0±3.8 years. The population included 26 patients with cleft palate with or without cleft lip who had previously undergone palatoplasty; 15 syndromic patients, including 11 with the diagnosis of velocardiofacial syndrome; 2 patients with VPI following adenotonsillectomy; and 7 patients with idiopathic VPI.

Summary statistics for the raw data are listed in Table 1. Interrater reliability ranged from 0.57 to 0.65 for all 3 components (right lateral wall movement plus left lateral wall movement; right palatal movement plus left palatal movement; and estimated percentage of gap size relative to the resting position). Among all raters, intrarater reliability was higher than interrater reliability for all rating categories, with intrarater reliability coefficient values ranging from 0.78 to 0.86 (Table 2). The faculty otolaryngologists rated segments more similarly to each other than did the fellows, and the least consistency was seen between the speech pathologists (Table 3). However, test-retest reliability for a given rater was similar among all rater types (Table 4). Reliability tended to be lower regarding qualitative characteristics than with quantitative measure-
Variability in degree of VP dysfunction has direct implications on patient management. The Golding-Kushner standardized scale was created to bring uniformity to the reporting of assessments of VP function. However, the interrater and intrarater reliability of the scale must be assessed if it is to serve its purpose as a tool to describe and communicate findings. Variability in reporting nasoendoscopic findings may have an impact on patient care, as surgical procedures for VPI management must be tailored to the patient's needs. Depending on the clinical situation, the endoscopist may need to convey the endoscopic findings to the managing surgeon. It is also important to understand the degree of variability in reporting endoscopic findings when comparing one examination with another. We need to understand the variability in describing VP function as we explore our institutional outcomes in VPI management. This will help us further refine patient selection criteria for various treatment options. Ultimately it will be important to understand variability in reporting of VP function between institutions to be able to generalize findings reported from a single institution and to pursue multi-institutional clinical protocols. The purpose of the present study was to understand variability of reporting nasoendoscopic findings within our institution.

D'Antonio et al. described flexible fiberoptic nasoendoscopy as a reliable method for assessing VP function and used a rating form that included both categorical and quantitative measurements. However, their study examined the ratings of groups of individuals and did not evaluate the interrater reliability of the lone rater, which may be a more realistic scenario at many institutions. The Golding-Kushner scale itself has not been validated.

Table 5. Interrater Reliability for Assessing the Presence of the Passavant Ridge, Midline Notch, Aberrant Pulsations, and Assignment of Gap Size Categories*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Interrater Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passavant ridge</td>
<td>0.44 (0.14-0.68)</td>
</tr>
<tr>
<td>Midline notch</td>
<td>0.57 (0.44-0.70)</td>
</tr>
<tr>
<td>Aberrant pulsations</td>
<td>0.58 (0.40-0.92)</td>
</tr>
<tr>
<td>Gap size category</td>
<td>0.53 (0.41-0.63)</td>
</tr>
</tbody>
</table>

*Data are reported as $\kappa$ coefficient (95% confidence interval).

Although reliability coefficients give us a measure by which to evaluate data regarding agreement between multiple observers, interpretation of these numbers should depend on the context and impact of variability. Landis and Koch attempted to provide criteria for interpreting reliability coefficients and created a scale in which 0 to 0.2 indicates “slight” agreement and greater than 0.8 indicates “almost perfect” agreement, with cutoffs for “fair,” “moderate,” and “substantial” agreement in between. If we judge our data by these criteria, there is a substantial degree of consistency between raters for quantitative measurements. However, the importance of the data lies not in such arbitrary labeling, but rather in that these figures provide us with a starting point, or benchmark, on which we must try to improve.

In our study, there was better correlation between the 2 attending otolaryngologists than there was between the fellows. The 2 attending raters have had extensive experience with VPI and routinely use the Golding-Kushner scale in making nasoendoscopic assessments. However, they had never compared or discussed their ratings. The discrepancy between the 2 fellows may have been related to the fact that one fellow was at the beginning of the training year while the other was at the end of clinical fellowship. The lowest degree of agreement was seen between the 2 speech pathologists. There was also a large difference in the experience levels of the 2 speech pathologists; the speech pathologist with less experience had the lowest intrarater and interrater reliability.

A decreased level of consistency for the categorical assessments was seen both between and within individual raters. This may be attributable, in part, to the infrequency of these characteristics, as the magnitude of the $\kappa$ statistic is sensitive to the frequency of occurrences. The lower level of consistency may possibly also have been due to rater fatigue, given that these assessments were made at the end of each rating form, after the quantitative assessments, which presumably required more focus and attention. There was a break of several seconds between each segment, and all raters described stopping the tape between some of the segments to regroup.

There were no consistent differences in reliability that correlated with segment length. We did not examine whether the amount of time each rater spent on each segment had an impact on reliability. Given the relatively lower consistency among raters and the better reliability among more experienced otolaryngologists, these data...
suggest that a training session to calibrate raters may be helpful in improving the interrater reliability. It is also possible that modifying the rating system would help to improve the consistency of reporting. Perhaps increments of 0.1 are not readily discernible units when rating endoscopic examinations. Consideration should be given to alternative rating scales.

Limitations of this study include the fact that there was no blinding to speech quality on the videotapes. It is therefore conceivable that the degree of speech compromise may have influenced the ratings. However, because the velopharynx is expected to close during swallowing, and conversely to remain open during generation of nasal sounds, it was important to include the audio information on the tapes. Another possible limitation was that the same videotape was viewed twice, with segments in the identical order. However, given the time lapse of weeks to months between the 2 ratings, it seems unlikely that recollection of previous ratings would have influenced interrater reliability.

The concept of tailoring interventions for VP dysfunction to an individual patient’s needs requires a method of assessing degree of VP closure that is reliable within an institution. To compare results from different centers, it is important to understand the patient population described. Ultimately, we will need a reliable method for reporting instrumental assessments as we design multicenter studies of surgical interventions for VPI.

In conclusion, the Golding-Kushner scale relies on subjective interpretation of qualitative visual information, which is naturally subject to variability. Our data suggest that this scale is a reasonably reliable tool for reporting endoscopic findings at our institution. However, it has also become clear that there is room for improvement in terms of reliability between raters. Future investigation may look at how training sessions affect interrater reliability.

Submitted for Publication: October 24, 2005; final revision received March 9, 2006; accepted March 22, 2006. Correspondence: Kathleen C. Y. Sie, MD, Children’s Hospital and Regional Medical Center, 4800 Sand Point Way NE, Seattle, WA 98105 (kathleen.sie@seattlechildrens.org).


Financial Disclosure: None reported.

Previous Presentation: This article was presented at the Annual Meeting of the American Society of Pediatric Otolaryngology; May 4-6, 2004; Phoenix, Ariz.

Acknowledgment: We gratefully acknowledge Judith Iwata for her assistance in data management, and Julie Dunlap, MS, CCC-SLP, and Linda Eblen, MA, CCC-SLP, for their participation in rating the nasendoscopic segments.

REFERENCES