Reliability of Flexible Fiberoptic Nasopharyngoscopy for Evaluation of Velopharyngeal Function in a Clinical Population

LINDA L. D’ANTONIO, PH.D.
JEFFREY L. MARSH, M.D.
MICHAEL A. PROVINCE, PH.D.
HARLAN R. MUNTZ, M.D.
CHRISTOPHER J. PHILLIPS, M.D.

Flexible fiberoptic nasopharyngoscopy (FFN) has become a popular clinical tool for evaluating velopharyngeal function. The literature contains numerous reports of FFN methodology and findings. However, there are few published reports that address clinical rating and reporting schemes or that evaluate viewer reliability. This study was designed to evaluate the reliability of visual perceptual ratings of FFN video images for assessing velopharyngeal structure and function in a clinical population. Ninety-five videotaped clinical evaluations were presented to and judged by three expert raters and nine novice raters from the fields of speech pathology, otolaryngology, and plastic surgery, using a standard rating form. A clinical rating scheme was used for quantifying perceptual judgments of velopharyngeal activity. Results suggest that videotaped FFN evaluations may be rated reliably, that expert raters working as a group are more reliable than novice raters working individually, and that the 125 evaluations presented without feedback are insufficient to improve the novices’ reliability. The combined auditory and visual perceptual evaluation inherent in FFN may be its most significant asset for both clinical and research applications.

KEY WORDS: nasopharyngoscopy, reliability, cleft palate, hypernasality, velopharyngeal insufficiency, speech, velopharyngeal function

Flexible fiberoptic nasopharyngoscopy allows direct transnasal observation of the anatomy and dynamic activity of the velopharyngeal sphincter. Such observations can be recorded for permanent documentation by coupling FFN to a videocamera with simultaneous audio recording. Numerous published reports discuss the advantages of FFN as a clinical method for evaluating velopharyngeal function during speech (Sawashima and Hirose, 1968; Sawashima and Ushijima, 1971; Miyazaki et al., 1975; Itoh et al., 1979; Matsuya et al., 1979; Bell-Berti, 1980; Croft et al., 1981; Ibuki et al., 1981; Karnell and Morris, 1985; D’Antonio et al., 1986). FFN may be of particular value for the cleft palate and craniofacial population for whom sequential management decisions concerning velopharyngeal function must be made. Direct visualization of velopharyngeal structure and function prior to treatment recommendations has become the standard of care in many clinical settings (Shprintzen et al., 1979; D’Antonio and Marsh 1987; Mackenzie-Stepner et al., 1987; D’Antonio et al., 1988). FFN is a valuable tool for direct visualization because it allows observation of the velopharyngeal valve during dynamic activity for a prolonged period with (1) minimal interference of the structures involved and (2) no radiation exposure.

Whereas most clinicians acknowledge the theoretic advantages of FFN and accept it as a valid technique for assessing velopharyngeal function, there are few published studies that have addressed the validity and reliability of endoscopic procedures. Endoscopic evaluation has high face validity (Dalston and Warren, 1985), and several reports have indicated that FFN has good construct validity when compared with radiologic assessments (Pigott and Makepeace, 1982; Sinclair et al., 1982; Ibuki et al., 1983; Karmel et al., 1983). Nonetheless, more research is necessary to confirm the validity of FFN as compared with other instrumental methods for use in answering specific questions regarding velopharyngeal structure and function. For
The video-taped clinical evaluations were reviewed to locate the velopharyngeal evaluations that were of acceptable quality and that had been conducted using a standard speech protocol, including the following productions: "pa, puppy, puffy, muffin, panmer, buy Bobby a puppy, Sissy sees the sun" and counting from one to ten. Evaluations were considered unacceptable if there were obscured anatomic landmarks or incomplete assessments of velopharyngeal function or if insufficient portions of the standard speech protocol were produced within a reasonable duration.

Ninety-five evaluations of 61 patients met the criteria for inclusion. Seventy evaluations were from patients who did not have a pharyngeal flap or other pharyngoplasty (non-flap), and 25 evaluations were from patients who had an intact pharyngeal flap (flap). There were 33 males and 28 females who ranged in age from 3 to 35 years. Sixty-one of the 95 segments were from initial evaluations, and 34 were from follow-up visits. The segments ranged in duration from 9 to 71 seconds, with an average of 45 seconds.

Thirty of the 95 segments were selected randomly for test-retest comparisons using a list of random numbers generated from the CRC Standard Math Tables, 25th edition (Beyer, 1978). The 30 retest segments and the 95 selected evaluations were combined and randomized using the same list of random numbers. A composite rating tape of the 125 randomly ordered segments was produced. Each segment was preceded by a dubbed audio announcement identifying the segment by number, beginning with "sample number 1" and ending with "sample number 125."

**Raters**

Twelve raters judged the 125 videotape segments included in the rating tape. Three raters worked together as a group to judge the rating tape and were considered the "expert group." This group was comprised of a speech pathologist, an otolaryngologist, and a plastic surgeon, all of whom had considerable experience collecting, interpreting, and viewing video-taped FFN evaluations of the velopharynx. The "expert group" arrived at a single rating because this was the usual clinical practice for the team at the hospital.

The remaining nine raters had little or no experience viewing or rating video-taped FFN evaluations. Three raters were practicing speech pathologists experienced in working with patients with cleft palate, three raters were residents or fellows in otolaryngology, and three raters were residents or fellows in plastic surgery. These nine novice raters judged the rating tape individually without consultation.

**Training**

Prior to judging the rating tape, the nine novice raters viewed and listened to a 30-minute training tape to ensure that all raters received the same instructions. They were familiarized with FFN view of the velopharynx and with a range of velopharyngeal closure patterns and speech samples that were representative of the types of clinical evaluations contained in the rating tape. Examples were shown of patients with and without pharyngeal flap. Anatomy and
function were discussed as they related to the rating tasks. The rating sheets were described in detail. No attempt was made to calibrate raters’ judgments of the training segments. In other words, raters were not instructed in specific quantitative ratings.

Raters were instructed to view the tape sequentially. Each segment could be viewed and listened to as many times as necessary, but the raters were instructed not to return to a previous segment. They were asked to complete the ratings in as few viewing sessions as possible and to keep a diary of the number of viewing sessions, duration of each viewing session, and segment numbers rated during each session. They received no feedback concerning their ratings and were instructed not to discuss their observations until the study was completed.

**Rating Forms**

Ten rating books were constructed—one for the expert group, which jointly judged the videotape segments as a single rater, and one for each of the nine individual, novice raters. Each book contained an individual rating form for each of the 125 segments to be judged. Two rating forms were used, one for the nonflap segments and one for the flap segments (Fig. 1).

The rating forms were designed to address whether closure was occurring for speech; to quantify movement of the velum, the right and left lateral pharyngeal walls, and the posterior pharyngeal wall during speech; and to assess relevant static anatomy such as velar contour and pharyngeal flap width. For the nonflap segments, a six-point equal interval scale was used to quantify movement of the velum, lateral walls, and posterior wall. Observers were instructed to rate movement of the specified velopharyngeal structures on a zero to five rating scale, where five indicated maximum movement of the velum to the posterior wall, of the lateral walls to midline, and of the posterior wall to the velum. A zero was used to indicate no movement of the structure. This system, along with the other questions, provided a description of the primary closure pattern for speech.

**A**

1. Can both lateral walls be visualized simultaneously?
   - Yes
   - No

2. Is the closure pattern symmetrical?
   - Yes
   - No

3. Is closure achieved for speech when expected?
   - Yes
   - No
   - Inconsistent

4. Velar contour
   - Hump
   - Noteh
   - Flat

Rate movement of the velum, lateral pharyngeal walls, and posterior pharyngeal wall on a 6 point scale with 0 representing no movement and 5 representing maximal movement.

**B**

1. Flap
   - wide
   - moderately wide
   - narrow

2. Patent
   - yes
   - no

3. Is closure achieved for speech when expected?
   - yes
   - no
   - Inconsistent

4. Is closure achieved for speech when expected?
   - yes
   - no
   - Inconsistent

5. Is one port larger?
   - right is larger
   - left is larger

6. Can both lateral walls be visualized simultaneously?
   - yes
   - no

7. Rate lateral wall motion on a 0-5 scale with 5 representing maximal movement to midline:

   **FIGURE 1** Rating forms used for A, the nonflap segments and B, flap segments. The rating forms were designed to address the following factors: whether velopharyngeal closure was occurring during speech; the quantification of the relative contributions of the velum, lateral pharyngeal walls, and posterior pharyngeal wall to closure; and the assessment of the static anatomy of relevant structures, such as the velum or a pharyngeal flap.
(Croft et al., 1981) and information about anatomic and functional symmetry.

For the flap segments, the width of the pharyngeal flap was rated as wide, moderately wide, or narrow. Binary alternatives (yes or no) were chosen for port patency and right and left lateral port closure for speech. Lateral wall motion was rated with the same six-point scale as that used for the nonflap segments. Anatomic and functional symmetry were also addressed. Velar movement in patients with flaps was not assessed because of an inability to create a reproducible grading scale.

In addition to the questions concerning velopharyngeal function for speech, the expert group was asked to rate the technical quality of each video segment as “good,” “fair,” or “poor.” When “fair” or “poor” were chosen, the expert group was asked to identify why the tape quality was judged as compromised. Following completion of the entire rating tape, each of the nine novice raters was asked to complete a questionnaire concerning his or her subjective impressions about the rating experience.

Statistical Analysis

To evaluate the reliability of the items on the rating forms, the Kappa statistic was employed (Fleiss, 1973). Kappa is similar to a correlation coefficient. It is a dimensionless measure of the amount of agreement among raters on a given item, with the maximum value of one indicating perfect agreement and the minimum value of zero indicating chance agreement only. Kappa corrects the fraction of observed agreement, $P_o$, for the fraction of agreement expected by chance, $P_c$, and is given by the following equation:

$$ K = \frac{P_o - P_c}{1 - P_c} $$

The Kappa statistic is an accepted measure of reliability for discrete, categoric ratings (yes/no, presence/absence, or the six-point scale as used in this investigation) (Cohen, 1968; Light, 1971; Fleiss, 1973). When an ordinal scale is used or when certain kinds of disagreement are considered less critical than other kinds, a weighted Kappa statistic is available (Cohen, 1968). To use this statistic, weights are assigned to every possible kind of disagreement to indicate the relative penalty. A weighted percent agreement is then calculated to arrive at the weighted Kappa. For the six-point scale items, for instance, we use the “natural” weighting function, which gives a penalty of zero for perfect agreement, a one for a disagreement of one grade, etc. The square of the distance between ratings (analogous to a “squared residual”) is used as the penalty. For example, the penalty function says that being off two grades one time is as bad as being off one grade four times. Cohen (1968) has shown that such a penalty function yields a weighted Kappa that is equivalent to the standard intraclass correlation coefficient, which treats the six-point rating scale as an interval one if the marginal probabilities are equal. Because there were more than two raters, we utilized the average of the pairwise calculated Kappas, which Light (1971) has shown to be a measure of the average overall agreement.

To test whether the Kappas are significantly different from zero (greater than chance agreement), the standard errors were computed to produce z-scores (Fleiss, 1973). These standard errors were also used to test the equality of the Kappas, in order to assess whether agreement was significantly different among raters, questions, or other relevant groups.

RESULTS

Consistency (Intrarater Reliability) of the Expert Group

Consistency is a measure of agreement between repetitive assessments by the same rater. Thirty of the 95 video segments were presented twice, randomly within the tape, so

### TABLE 1 Consistency (Intrarater Reliability) for Expert Group

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Weighted % Agree</th>
<th>Weighted Kappa</th>
<th>s.e.</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NONFLAP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both lateral walls viewed?</td>
<td>21</td>
<td>90</td>
<td>0.738</td>
<td>0.289</td>
<td>0.0054</td>
</tr>
<tr>
<td>Closure pattern symmetry*</td>
<td>16</td>
<td>100</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closure for speech*</td>
<td>21</td>
<td>85</td>
<td>0.687</td>
<td>0.182</td>
<td>0.0001</td>
</tr>
<tr>
<td>Velar contour</td>
<td>21</td>
<td>93</td>
<td>0.774</td>
<td>0.238</td>
<td>0.0006</td>
</tr>
<tr>
<td>Right wall motion*</td>
<td>16</td>
<td>96</td>
<td>0.650</td>
<td>0.227</td>
<td>0.0021</td>
</tr>
<tr>
<td>Left wall motion*</td>
<td>19</td>
<td>96</td>
<td>0.723</td>
<td>0.193</td>
<td>0.0001</td>
</tr>
<tr>
<td>Posterior wall motion*</td>
<td>21</td>
<td>99</td>
<td>0.725</td>
<td>0.359</td>
<td>0.0218</td>
</tr>
<tr>
<td>Velum motion*</td>
<td>19</td>
<td>94</td>
<td>0.496</td>
<td>0.209</td>
<td>0.0088</td>
</tr>
<tr>
<td><strong>FLAP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width of flap</td>
<td>9</td>
<td>94</td>
<td>0.757</td>
<td>0.398</td>
<td>0.0286</td>
</tr>
<tr>
<td>Right port patent?</td>
<td>9</td>
<td>100</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left port patent?</td>
<td>9</td>
<td>100</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right port closure†</td>
<td>9</td>
<td>83</td>
<td>0.658</td>
<td>0.289</td>
<td>0.0114</td>
</tr>
<tr>
<td>Left port closure†</td>
<td>9</td>
<td>83</td>
<td>0.649</td>
<td>0.296</td>
<td>0.0142</td>
</tr>
<tr>
<td>Largest port</td>
<td>3</td>
<td>100</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both lateral walls viewed?</td>
<td>8</td>
<td>62</td>
<td>0.143</td>
<td>0.401</td>
<td>0.3608</td>
</tr>
<tr>
<td>Right wall motion†</td>
<td>8</td>
<td>96</td>
<td>0.500</td>
<td>0.312</td>
<td>0.0544</td>
</tr>
<tr>
<td>Left wall motion†</td>
<td>9</td>
<td>96</td>
<td>0.625</td>
<td>0.258</td>
<td>0.0077</td>
</tr>
</tbody>
</table>

Consistency (intrarater reliability) data for the expert group for each variable for the nonflap and flap segments. Variables that were considered to characterize velopharyngeal function are designated with an * for the nonflap segments and with a † for the flap segments. There was no significant difference in consistency between the nonflap and flap segments ($p = 0.032$).
that intrarater reliability could be assessed. Twenty-one of
the repeated segments were for nonflap patients and nine
repeat segments were for flap patients. The agreement be-
tween the initial rating and the repeat rating provides infor-
mation about the consistency of an individual rater.

For the expert group, rating as a consensus unit, there
was good agreement for all nonflap variables and for all but
one flap variable (Table 1). For the nonflap segments, there
was 100 percent consistency for question two, "Is the clo-
sure pattern symmetrical?". The lowest consistency score
for the nonflap segments was for question three, "Is closure
achieved for speech when expected?" (85 percent weighted
agreement, weighted Kappa = 0.68). For the flap segments,
the three questions regarding patency and size of the ports
had 100 percent consistency between first and second rat-
ings for the expert group. The lowest agreement for the flap
segments occurred for the question, "Can both lateral walls
be visualized simultaneously?" (62 percent weighted agree-
ment, weighted Kappa = 0.14).

Examination of comparable questions in the nonflap and
flap groups in Table 1 documents considerable agreement in
the Kappas between the two sets of patients. This suggests
that flap segments were not significantly "harder" to eval-
uate than were nonflap segments (p = 0.4361). The one ex-
ception to this generalization was the question, "Can both lateral walls
be viewed simultaneously?" This question had
a Kappa of 0.74±0.29 for the nonflap segments but only
0.14±0.40 for the flap segments (p = 0.032).

Consistency (Intrarater Reliability) of the
Novice Rater

The initial questions of interest concerning the nine in-
dividual novice raters were whether they were consistent,
and if so, whether they differed in consistency from the
expert group. Whereas the consistency of the expert group
was high for almost all questions, with a Kappa value be-
tween 0.50 and 1.00 (see Table 1), the novice group was
less consistent, with intrarater Kappas between 0.30 and
0.80 (Table 2). For some questions, the mean Kappas were
lower than for others, which indicates that some raters had
poorer consistency measures for those questions. However,
there was no readily evident pattern in the lower scores.
Although some of the mean Kappa values appear low, they
are all, nonetheless, significantly different from chance
agreement (i.e., K = 0). Therefore, the overall Kappas in-
dicate that, on the average, novice raters were consistent
in their ratings and mean consistency scores were significantly
greater than chance (p = 0.0380).

For the nonflap segments, consistency was lowest and
differed significantly from the expert group most often for
the three following variables: posterior wall motion, velar
contour, and the question, "Can both lateral walls be visu-
alyzed simultaneously?". For the flap segments, consist-
ency scores for right port closure and left wall motion were
low, with two of the novice raters differing significantly
from the expert group consistency scores.

Group Differences in Consistency

The mean weighted Kappa scores for each group of raters
were averaged for the six variables that characterize veloc-
pharyngeal function. The average Kappa for the expert
group was 0.78. The Kappa values for the novice group
ranged from 0.51 to 0.59 (speech, 0.59; plastic surgery,
0.53; ENT, 0.51). Comparing the mean consistency ratings
of the three novice professional groups with those of the
expert group (Fig. 2), there was no statisticallysignificant
difference in consistency between the expert group and
the speech pathologists (p = 0.2739). However, the mean con-
sistency measures for the novice otolaryngologists and the
novice plastic surgeons were significantlylower than those
for the expert group (ENT, p = 0.0001; plastic surgery,
p = 0.0036).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean Weighted Kappa</th>
<th>Minimal Kappa</th>
<th>Maximal Kappa</th>
<th>No. Differing Significantly from Expert Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONFLAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both lateral walls viewed?</td>
<td>0.320</td>
<td>-0.061</td>
<td>0.608</td>
<td>2</td>
</tr>
<tr>
<td>Closure pattern symmetry</td>
<td>0.520</td>
<td>0.250</td>
<td>0.889</td>
<td>1</td>
</tr>
<tr>
<td>Closure for speech</td>
<td>0.637</td>
<td>0.269</td>
<td>0.917</td>
<td>3</td>
</tr>
<tr>
<td>Velar contour</td>
<td>0.377</td>
<td>-0.057</td>
<td>0.639</td>
<td>0</td>
</tr>
<tr>
<td>Right wall motion</td>
<td>0.565</td>
<td>0.398</td>
<td>0.828</td>
<td>0</td>
</tr>
<tr>
<td>Left wall motion</td>
<td>0.514</td>
<td>0.384</td>
<td>0.838</td>
<td>2</td>
</tr>
<tr>
<td>Posterior wall motion</td>
<td>0.348</td>
<td>-0.089</td>
<td>0.954</td>
<td>1</td>
</tr>
<tr>
<td>Velum motion</td>
<td>0.470</td>
<td>-0.037</td>
<td>0.896</td>
<td></td>
</tr>
<tr>
<td>FLAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width of flap</td>
<td>0.434</td>
<td>0.143</td>
<td>0.727</td>
<td>0</td>
</tr>
<tr>
<td>Right port patent</td>
<td>0.667</td>
<td>0.000</td>
<td>1.000</td>
<td>2</td>
</tr>
<tr>
<td>Left port patent</td>
<td>0.889</td>
<td>0.000</td>
<td>1.000</td>
<td>1</td>
</tr>
<tr>
<td>Right port closure</td>
<td>0.378</td>
<td>-0.455</td>
<td>0.883</td>
<td>0</td>
</tr>
<tr>
<td>Left port closure</td>
<td>0.512</td>
<td>-0.500</td>
<td>0.889</td>
<td>2</td>
</tr>
<tr>
<td>Largest port</td>
<td>0.597</td>
<td>0.000</td>
<td>1.000</td>
<td>1</td>
</tr>
<tr>
<td>Both lateral walls viewed?</td>
<td>0.586</td>
<td>0.372</td>
<td>1.000</td>
<td>0</td>
</tr>
<tr>
<td>Right wall motion</td>
<td>0.364</td>
<td>0.118</td>
<td>0.680</td>
<td>0</td>
</tr>
<tr>
<td>Left wall motion</td>
<td>0.363</td>
<td>0.022</td>
<td>0.700</td>
<td>2</td>
</tr>
</tbody>
</table>

* Tests of statistical differences between the expert group and novice raters could not be calculated when the expert consistency was 100 percent.
Interrater Reliability

Interrater reliability is a measure of agreement between raters. Interrater reliability compares a novice's evaluation of a segment to the expert group evaluation of that same segment. Nine segments were identified by the expert group at the time of their evaluation as being of "poor technical quality" (excessive light reflection, inadequate illumination, or excess mucus) and were therefore difficult to rate. Since the interrater reliability for these segments would be expected to be lower than for the rest of the segments for technical reasons, they were excluded from calculations of interrater Kappas. The interrater Kappas, therefore, should be interpreted as measuring the agreement of novices to experts for those tape segments that the experts felt to be "evaluable." That is, any error in the novice's evaluation cannot be attributed to the technical quality of the segments.

The data were analyzed to determine how each novice rater compared with the expert group for ratings of the six variables selected as characterizing velopharyngeal function. Data from the three individuals within each professional discipline novice group were pooled. Each of these groups was then compared with the expert group. These comparisons generated mean interrater Kappas (Fig. 3). The average Kappas were between 0.30 and 0.40, with some difference among professions (speech pathology, 0.40; plastic surgery, 0.39; ENT, 0.37). Since all Kappa values were significantly different from zero, which indicated greater than chance agreement with the expert group, it can be concluded that novices are able to evaluate FFN videos with reasonable reliability. However, the average ratings by the novices did differ from those of the experts. One can assume that the best the novices could do would be to agree with the experts to the same degree that the experts agree with themselves (on test-retest). Thus, if the novices were "just as good" as an expert group, then their interrater Kappas would be nearly as large as the expert's intrarater reliability, which is in the 0.60 to 0.80 range. In reality, the interrater reliability between the novices and experts is only about half that "maximum" value. This gives some indication of the magnitude for improvement that remains in novice rating.

Group Differences in Interrater Reliability

The data for each of the novice rater groups were analyzed to determine whether there was a statistical difference between each of the three professional groups, respectively, in the magnitude of their agreement with the expert group. The mean Kappa was calculated for the composite three raters in each professional group for the six variables characterizing velopharyngeal function (see Fig. 3). There was no significant difference between the speech pathologists and plastic surgeons (p = 0.1356) in their agreement with the expert group. However, the otolaryngologists had significantly lower agreement with the expert group when compared with the speech pathologists (p = 0.0003) or the plastic surgeons (p = 0.0336).

Changes in Interrater Reliability Over Time

The present study was not designed specifically to assess the effects of learning on reliability. However, the rating tape was divided into quartiles to assess changes in agree-
ment between the novice raters and the expert group over time (i.e., changes in interrater reliability over time). There were no differences between the first three quartiles (Fig. 4). Novice raters began the study with a given level of agreement with the expert group (0.40 range), and this level of agreement did not change until the last quartile, and their level of agreement decreased (0.34).

Postexperiment Questionnaire

Following completion of the experiment, the novice raters completed a questionnaire and were interviewed concerning their experience with the study. The inquiries were of the four following types: anatomic, rating scale, means of perception, and education. The raters varied in their responses as to whether differences in anatomy affected the ease of rating. For example, five raters reported no difference in difficulty in rating nonflap versus flap segments. Two raters found nonflap segments easier, and two raters preferred flap segments. Similarly, there was diversity among raters when asked whether any questions were more difficult than others. Three raters found velar contour most difficult, and two raters were uncomfortable with assessing port patency and port size. When asked if the zero to six interval scale was practical for the individual rater, two-thirds of the novice raters responded positively. Three raters expressed dissatisfaction with a zero to six scale, and two of these raters expressed a preference for a zero to three scale.

One question resulted in a clear difference among the novice professional groups. Most raters reported that they relied on both visual and auditory information. One plastic surgeon, however, reported sole reliance on visual information. All three speech pathologists reported heavy reliance on auditory perceptual information over the visual modality.

Several follow-up questions focused on whether raters felt qualified to make judgments of FFN for clinical management. One rater felt qualified both before the experiment and after. Two raters felt the experiment had provided them with enough practice to make similar clinical judgments. Two-thirds of the raters believed they were more experienced following the experiment than before. However, they did not consider themselves adequately qualified to make clinical judgments. There was a recurring theme among the raters' responses, in that all raters expressed a desire for feedback or working in a group. When asked how they would train others to do a similar task, they reported an overwhelming preference for calibration tapes. Most raters suggested training tapes with feedback and anchoring of perceptual judgments.

**DISCUSSION**

An important consideration concerning FFN as it is used in clinical practice is that its processing relies on examiner interpretation of visual and auditory information by one or more clinicians. As with any perceptual evaluation, estimates of inter- and intrarater reliability are important. If a methodology is to be utilized clinically or for clinical research, there should be data available concerning the reliability of judgments and the process of training individuals to be competent clinical users. The purpose of this study was (1) to address some questions of reliability that would be most likely to affect clinical use of FFN in evaluating velopharyngeal function and (2) to compare expert and novice raters in order to address some issues regarding training.

The findings of this study suggest that experienced clinicians working together as a group can make highly consistent judgments concerning velopharyngeal structure and function during speech from FFN videotapes. This study did not address the validity of endoscopic assessments, which has been addressed previously (Zwitman et al, 1976; Pigott and Makepeace, 1982; Sinclair et al, 1982; Ibuki et al, 1983; Karnell et al, 1983). Although there was no difference in rater consistency between nonflap and flap segments, the validity of judgments of FFN views of patients with pharyngeal flap has not been established adequately and should be studied further.

Analysis of consistency scores for the various questions showed individual differences among the raters in their performance on specific questions. There was no pattern that suggested that some items were intrinsically more difficult than others. These results suggest that, before engaging in clinical decision making based in interpretation of FFN videos, a given judge might benefit from knowing which variables he or she can expect to judge reliably. Furthermore, it is likely that training could be helpful in overcoming a rater's specific weaknesses once they are known.

This study compared the performance of experts working as a group with the performance of novices working alone. No attempt was made to isolate the effects of group performance compared with individual performance nor to isolate experts working individually compared with novices working alone. These are important comparisons that deserve further investigation, but were beyond the scope of the present study. Rather, this experiment was designed to compare the extremes of clinical practice.
Comparison of the expert group and novice raters shows that experienced clinicians working together as a group are more consistent than are novice raters working alone. The design of the study did not allow for a comparison of experts working alone versus experts working in a group. However, our own clinical experience suggests that experts working in a group are more reliable than they would be in isolation. While this observation is not surprising, it is clinically relevant. Dalston and Warren (1985) have discussed possible sources of error in auditory perceptual evaluations. They point out that, although listener evaluations have high face validity, they are frequently unreliable. Reliability, however, may be improved dramatically by training and by pooling judgments from multiple listeners. It is likely that the same logic applies to FFN evaluations.

Inspection of mean consistency and agreement data showed some performance differences among the novice professional groups. In this study, the speech pathologists were, on the average, more consistent than the other two novice groups and more similar to the experts than to the other novice groups. The results showing statistical differences between the professional groups should be interpreted cautiously because of the small sample size.

However, postexperiment interviews documented other group differences. Speech pathologists spent more time viewing the videos than did their physician colleagues, which may account for their performance. An additional difference among the groups, which emerged from the post-experiment interviews, may have important implications for clinical use and training. Specifically, all of the speech pathologists reported a high dependence on auditory information compared with visual information. Two of the speech pathologists "confessed" to listening first and then reviewing the tape for visual information. The subjective reports of the speech pathologists and their reliable performance may be interpreted as support for the notion that combined recordings are superior to single modality evaluations. FFN provides simultaneous visual and auditory information. Much emphasis has been placed on the visual data obtained from FFN assessments.

Our results suggest that training should teach new raters to use both auditory and visual information in making their evaluations. Additionally, it is possible that group consensus ratings, which are arrived at by experts who differ in preference for visual versus auditory information, may be superior to ratings by individuals who are primarily experienced in one modality.

Other data that provide insights concerning training are found in the assessment of change over time. As described earlier, the novice raters began the study with a given level of agreement with the expert group. This level of agreement was significantly different from chance but was still in the "moderate" range, with room for improvement. In the last quarter of the experiment, the novice ratings diverged significantly in a negative direction from the expert group when compared with the previous three-quarters. Fatigue, boredom, or overload could account for the pronounced decrease in performance. However, a more interesting finding is that no improvement in performance occurred over time. This finding implies that exposure to 125 video FFN segments without discussion or feedback is insufficient to improve agreement between the novice raters and the experts. These data support the raters' responses to the post-experiment questionnaire, which stated that they needed feedback or calibration. These data also raise the question of whether training and anchoring are the best means of providing necessary feedback or whether group judgments would effectively increase the performance of the novices as compared with that of the expert group. Regardless of whether one or both methods would increase performance, the fact that exposure to 125 segments in isolation was insufficient to increase performance is relevant for centers considering the addition of FFN to their diagnostic battery.

Acknowledgement. We are indebted to Karen Lask-McDonagh for manuscript preparation.

REFERENCES


MACKENZIE-STEPNER K, WITZEL MA, STRINGER DA, LINDSAY WK, MUNRO IR, HIGHERS H. (1987). Abnormal carotid arteries in the velo-
Commentary

It is this reviewer’s privilege to comment on the excellent article by D’Antonio et al. This manuscript is an extremely important work. The authors are to be commended on the detail in which they evaluated the reliability of nasopharyngoscopy in the assessment of velopharyngeal function. Their finding that experienced clinicians who work together as a group are highly consistent in their accuracy concerning the assessment of velopharyngeal function is important. It appears that one of the major points of this paper is that it is very important for experts to have feedback and discussion among themselves. The authors accurately point out that this particular study did not attempt to isolate individual performance versus group performance, nor could they comment on a comparison with experts working individually versus with novices working in a similar situation. One must wonder whether the novices working as a group would have improved their accuracy, and if so, whether it would have been close to the accuracy of the expert group.

It is also important to note that their results suggest (but, I do not feel, conclusively prove) that trainee orientation should teach new raters to use both auditory and visual information. A possible future study that would be of value would be to evaluate the use of visual input only, auditory input only, and the combination of visual and auditory input. This commentator could conceive of a study design that would accomplish this goal.

Turning our attention to some specific items, the authors comment that “direct visualization of velopharyngeal structure and function prior to treatment recommendations has become a standard of care in many clinical settings.” I would like to point out that I do not feel that this has become the standard of care, in spite of our wishes that it should be. This type of evaluation may be impractical in many settings because it takes not only a relatively large amount of time from two or three specialties, but it also requires a significant amount of equipment and storage facilities, which may not be feasible in many practices. It is my feeling that tertiary centers (such as cleft palate and craniofacial centers) must provide this type of evaluation; however, it would be unreasonable to expect this to be the standard among all practitioners.

The authors discuss the “face validity” and the “construct validity” of fiberoptic nasopharyngoscopy, but they do not define these terms; it would have been helpful had they defined the terms in their article. They also point out that little is known concerning how the estimates of velopharyngeal orifices using nasopharyngoscopy compare with those of other modalities. It would be interesting to correlate pressure flow measurements with estimates of function, as described in this paper.

In discussing the single rating by the “expert group,” the authors may have biased or skewed the evaluations. A range of rating would be helpful to see whether any ratings were significantly influenced by other members of the “group.” Another piece of information that would be useful to evaluate the practicality of such a method in the average provider’s practice would be to examine the range and the mean time per patient evaluation. In the average practitioner setting, such information would be vital, because it may be difficult to get two or three busy healthcare providers together for a prolonged period of time during which they could evaluate a number of videotapes.

The authors are to be commended on the evaluation of the intrarater reliability (consistency) and the interrater reliability. This is extremely important information, and the methods utilized for reliability were satisfactory. The removal of nine segments judged to be of “poor technical quality” is very important, since all tapes should be of sufficient quality to be “evaluable.” There is some concern when the authors “pool” data from three individuals within each professional discipline novice group. This information could be extremely variable depending on how the combinations of individuals were placed. In other words, three accurate novices could be put together, three of the most accurate pooled, or any combination thereof. The various combinations would affect the data to be analyzed.

In summary, I feel that the purpose of this study was well defined and that the authors satisfactorily accomplished their goals. They have also raised questions that could be the basis for further investigative efforts.

Charles F. Koopman, Jr, M.D.
Associate Professor of Otolaryngology
University of Michigan