Dysphagia 13:223–227 (1998)

Dysphagia

A Study of Interrater Reliability when Using Videofluoroscopy as an Assessment of Swallowing

Amanda Scott, B App Sci,[2] Alison Perry PhD, MRCSLT,[1] and John Bench, BSc (Hons), BA (Hons), PhD, MAPsS, FBPsS, CPsychol, FAudSA (CC)[1]

[1]School of Human Communication Sciences, La Trobe University, Bundoora, Victoria, Australia; and [2]School of Human Communication Sciences, La Trobe University and Bethlehem Hospital, South Caulfield, Victoria, Australia

Abstract. Videofluoroscopic assessment of swallowing is widely used in clinical settings. The interpretation of such assessments depends on subjective visual judgments but the reliability of these judgments has been poorly researched. This study measured interrater reliability of judgments, made by speech pathologists, of videofluoroscopic images of subjects swallowing liquid and semisolid boluses. A 5-point rating scale was used in three conditions: individually after careful reading; together with other speech pathologists in group discussion; and individually after the group discussion. Analysis of the ratings for the three conditions revealed that the level of agreement among raters was generally higher for semisolid swallows than for liquid swallows. The highest levels of agreement occurred for ratings made after group discussions. The levels of agreement were lowest when raters worked alone, relying only on reading the scale. Individual rating after group discussion resulted in higher levels of agreement than sole reliance on reading the scale. Factors influencing the levels of interrater agreement, including the timing of observations, bolus consistency, the quality of the image, and the complexity of the task, are discussed.

Key words: Dysphagia — Videofluoroscopy — Interrater reliability — Deglutition — Deglutition disorders.

Swallowing is a complex activity composed of a series of rapid, integrated movements, few of which can be observed directly. The development of videofluoroscopic techniques has enabled the swallowing process to be visualized [1] and has enhanced the understanding of swallowing. Following the work of Logemann et al., videofluoroscopy clinics (usually involving radiologists in conjunction with speech pathologists) have proliferated as a means of diagnosing, assessing, and managing patients with swallowing disorders. However, the clinical interpretation of videofluoroscopic images depends on subjective judgments and there has been little systematic attempt to validate these. Given the importance of some management decisions based on the interpretation of videofluoroscopic swallowing assessments, such as the permanent cessation of oral intake; surgical intervention (e.g., cricopharyngeal myotomy); a reliable, methodologically sound means of interpreting videofluoroscopic assessments is needed.

Various approaches to standardizing the interpretation of videofluoroscopic assessment of swallowing have been made. For example, the Videofluoroscopic Worksheet in Logemann's [2] ''Manual for the Videofluoroscopic Study of Swallowing'' provides a checklist of possible signs to be matched with clinicians' subjective observations. This approach allows clinicians to make binary decisions about a range of variables. Its strength is in the descriptions of relevant observations based on expert clinical experience. However, this approach does not allow degrees of change over time to be measured in individuals. Further, the ability of clinicians to accurately recognize the radiographic signs outlined in the worksheet has not been tested.

Ranking methods have been used for the levels of function of individual factors during videofluoroscopic assessment of swallowing. The ''Dysphagia Profile'' [3], developed at Charing Cross Hospital, London, is a means of ranking observations using a 5-point scale. This profile measures eight aspects of swallowing: lip function, tongue function and bolus control; soft palate function;

Correspondence to: Amanda Scott, B App. Sci., Speech Pathology Department, Bethlehem Hospital, 476 Kooyong Road, Caulfield South, Victoria 3162, Australia

triggering of the swallow reflex; pharyngeal motility, aspiration; cricopharyngeal function; and upper esophageal obstruction. The profile provides a protocol for graded observations but again lacks published validity and reliability information, including comparisons with normative data. Hence, the assumptions made when using the profile are of unknown validity and reliability.

The issue of interrater variability in assessing dynamic images of swallowing has been addressed by Ekberg et al. [4], Gibson and Phyland [5], and Wilcox et al. [6]. These studies all demonstrated the variability in interpretation of videofluoroscopic images of swallowing. Ekberg et al. found that higher levels of agreement were obtained when unequivocal functions such as ''normal pharyngeal function'' (0.83 Kappa coefficient) or ''aspiration of barium into the trachea'' (0.70 Kappa coefficient) were assessed. However, poorer concordance was noted when less definite levels of dysfunction were measured, for example, ''decreased pharyngeal constriction'' (0.22 Kappa coefficient) or ''the presence of a cricopharyngeal impression of less than 50%'' (0.40 Kappa coefficient).

Wilcox et al. [6] reported the results of speech pathologists making binary judgments of videofluoroscopic recordings of swallowing. Overall, a level of agreement of 85.33% was obtained for the total of 256 judgments made in this study. Disagreements accounted for 14.66% (n = 44) of observations, with the majority of disagreements related to pharyngeal deficits (n = 26). Over half of these disagreements specifically related to pooling in the valleculae and pharynx (n = 16).

Gibson and Phyland's [5] study of 4 speech pathologists rating 20 videofluoroscopic swallows from 8 subjects with dysphagia (each judgment being performed twice), reported high interrater reliability measures for the oral and pharyngeal phase transit times and for the number of swallows required to clear the pharynx of the bolus. There were also good levels of agreement in recognizing the position of the bolus at the initiation of the swallow. However, as with the earlier reported studies, interrater reliability was worse for the measurement of pooling in the valleculae. Interestingly, a good level of agreement was reached rating this function on the first assessment; but the amount of agreement on the second assessment was poor. The authors suggested that this discrepancy was due in part to the use of a 3-point scale for the rating of vallecular pooling, rather than the simpler binary decisions used for rating the other factors.

From a review of the literature it is evident that no appropriate level of interrater agreement among clinicians using videofluoroscopic assessment of swallowing has been achieved to date. This paper aims to clarify the process of interpreting videofluoroscopy and to define the areas of inconsistency and potential misinterpretation.

Materials and Methods

Subjects

Nine speech pathologists were recruited from a professional interest group specializing in dysphagia. Two raters considered themselves to be ''very experienced'' in the interpretation of videofluoroscopic assessment of swallowing; 5 reported that they were ''moderately experienced''; and 2 had ''minimal experience'' in the area of videofluoroscopy.

The Task

Each rater was given a videotape containing fluoroscopic recordings of swallowing by 3 individuals (2 patients with dysphagia due to motor neurone disease and 1 volunteer without dysphagia). Each swallowed six liquid and six semisolid barium mixture boluses. The swallows were viewed in lateral perspective, the fluoroscopic image focused on the oral cavity for the first three swallows and on the pharynx for the second set of three swallows. Each sample was numbered and the order of swallows on the tape was randomly assigned.

The Scale

The 5-point rating scale used in this study was based on the Charing Cross Hospital dysphagia profile [3]. The scale contains 11 subtests, 3 of which assessed the oral phase: lip, tongue, and jaw function. The remaining 8 subtests assessed the pharyngeal phase: velar, hyoid, pharyngeal wall, and cricopharyngeal movement, pooling of barium in the valleculae and pyriform sinuses, and the presence of aspiration. A description of the expected observations is outlined at each level of a 5-point scale, with level 1 representing optimum function and levels 2, 3, 4, and 5 representing decreasing levels of function (see Table 1).

Procedure

The subjects were given the tape containing videofluoroscopic swallowing assessments from the 3 patients, a copy of the Scale, and three rating forms. They were informed that . . . ''the recordings of swallows might be chosen to exhibit an example of dysphagia or not.'' No other information was given.

Condition 1. ''Individual Use of the Scale without Experience in its Use. Raters were instructed to ''read the Scale carefully and rate the first subject on the tape using the Scale. Record, in order of priority, any comments or difficulties experienced.'' Two weeks were allowed for the completion of Condition 1.

Condition 2. ''Individual use of the Scale, with conferring.'' The raters met once with the investigator (AS) in three groups (n = 2, 2, 5). The first part of these meetings was devoted to a discussion of issues arising from the raters' use of the Scale during Condition 1. Their comments were noted. Each group then rated the second swallow on the tape, with discussion between the raters. During this part of the study the investigator answered specific questions about the Scale and/or task and facilitated discussion between raters. She did not disclose her own rating until the group reached a consensus.

Condition 3. ''Individual Use of the Scale after Experience in its Use.'' Within the remaining 2 weeks the raters independently assessed the third subject using the same Scale.

All raters participated in Conditions 1 and 2, however, for Condition 3, only 7 speech pathologists remained in the study.

Table 1. Example of rating scale

Tongue function—levels of function
1. Bolus is propelled competently into pharynx in a smooth, uninterrupted wave-like motion.
2. Bolus is propelled competently, as in 1, but divided into two sections, and/or bolus propelled into pharynx, as in 1, but slowly, and/or tentative initial tongue movements prior to propulsion of the bolus into the pharynx, and/or disruption of the wave-like motion during propulsion with 2–5 pushes required to transport the bolus into the pharynx, and/or oral structures become coated with the barium mixture.
3. Propulsion of the bolus into pharynx is disrupted with six or more pushes required to propel the bolus into the pharynx, and/or bolus divides into two sections in the presence of impaired tongue movement, and/or a small amount of the bolus remains in the oral cavity after the swallow.
4. Bolus is propelled into pharynx in piecemeal manner (more than three sections), and/or tongue movement is slow and reduced in range, and/or about half the bolus pools in the oral cavity after swallow.
5. Minimal movement of tongue, and/or subject uses finger or spoon to push bolus back in mouth, and/or most of the bolus remains in the oral cavity after swallow.

Results

Raters' scores for each of the 11 subtests, under each of the three conditions, were compared for liquid and semisolid bolus presentations with Spearman's Rho. To allow for variations in the distributions of the rho-values and to facilitate later analysis, the rho-values were converted to Fisher's Z_r scores. A mean Z_r value was then calculated for each subtest (Tables 2 and 3).

The agreement for semisolid bolus presentations was generally higher and attained higher levels of statistical significance than the agreement for liquid swallows ($t_{32} = 2.21$, $p < 0.05$). As expected, the levels of agreement were lowest for Condition 1, with Z_r scores for 3 of the 11 subtests for semisolid swallows and only 1 of 11 subtests for liquid swallows, reaching the 0.05 level of significance or better. The highest levels of agreement occurred under Condition 2, with Z_r scores for all of the 11 subtests for both semisolid swallows and liquid swallows, reaching levels of 0.05 or greater. Eight of the 11 subtests for semisolid swallows and 4 for liquid swallows reached levels of significance of 0.005. Variations in levels of agreement in Condition 2 were mainly due to differences in interpretation between the groups of raters. The tendency to obtain higher Z_r scores and higher levels of significance for semisolid bolus presentations was maintained under Condition 3. Six subtests remained at the 0.05 or higher levels of significance compared with five for the liquid bolus presentations. The subtests measuring labial function, lingual function, and hyoid bone elevation during semisolid swallows, and labial function and the presence of pooling in the pyriform sinuses during liquid swallows, went against the overall tendency for higher levels of agreement to occur in Condition 3 compared with Condition 1.

Table 2. Mean Z_r scores for semisolid bolus presentations

Subtest	Condition 1 (9 Raters)	Condition 2 (9 Raters)	Condition 3 (7 Raters)
Labial function	1.34	3.00[c]	1.15
Lingual function	1.49	1.92[a]	1.40
Jaw function	1.02	3.00[c]	2.30[b]
Velar function	1.09	1.92[a]	3.00[c]
Swallow reflex	1.28	3.00[c]	1.37
Hyoid elevation	2.35[b]	3.00[c]	1.70[a]
Pooling in valleculae	1.25	3.00[c]	1.52
Pooling in pyriform sinuses	2.03[a]	3.00[c]	3.00[c]
Aspiration	2.18[a]	3.00[c]	3.00[c]
Pharyngeal wall function	1.10	1.92[a]	2.35[b]
Cricopharyngeal function	1.19	3.00[c]	1.38
Mean	1.48	2.71[c]	2.02[a]
Standard deviation	0.48	0.50	0.73

[a]$p < 0.05$; [b]$p < 0.01$; [c]$p < 0.005$.

Table 3. Mean Z_r scores for liquid bolus presentations

Subtest	Condition 1 (9 raters)	Condition 2 (9 raters)	Condition 3 (7 raters)
Labial function	3.00[c]	2.12[a]	1.70[a]
Lingual function	1.13	1.79[a]	1.77[a]
Jaw function	1.94	2.24[a]	2.35[b]
Velar function	1.05	1.70[a]	1.47
Swallow reflex	0.81	1.70[a]	1.47
Hyoid elevation	0.76	3.00[c]	1.70[a]
Pooling in valleculae	0.96	2.24[a]	1.28
Pooling in pyriform sinuses	1.13	3.00[c]	0.75
Aspiration	1.17	1.79[a]	3.00[c]
Pharyngeal wall function	0.62	3.00[c]	1.48
Cricopharyngeal function	1.32	3.00[c]	1.48
Mean	1.26	2.34[b]	1.81[a]
Standard deviation	0.67	0.55	0.65

[a]$p < 0.05$; [b]$p < 0.01$; [c]$p < 0.005$.

A factorial analysis of variance was performed on the Z_r scores for the 11 subtests under the three Conditions, with separate analyses of liquid and semisolid bolus presentations. Differences in the Z_r scores obtained within the three Conditions were significant at the 0.001 level ($F_{2,20} = 39.85$). A post hoc comparison of the means for the factor 'Condition,' using Tukey's Honestly Significant Difference (HSD) Test (Tukey) [7] indicated that scores obtained under Condition 2 were significantly different from the scores obtained for Conditions 1 at the 0.05 level, but not for Condition 3 (Table 4).

The Z_r scores obtained for liquid and semisolid boluses showed significantly better interrater agreement for semisolid swallows at the 0.05 level ($t_{32} = 2.13$, $p <$

Table 4. Differences between the mean Z_r of each condition

	Condition 1	Condition 2	Condition 3
Condition 1	0	2.25[a]	1.22
Condition 2		0	1.03
Condition 3			0

[a]HSD (Honestly significant difference) = 2.06, significant at 0.05 level.

0.05). However, there were no significant differences between the Z_r scores for the 11 subtests ($F_{10,20} = 1.67$).

The effect of the higher levels of interrater agreement for semisolid swallows compared with liquid swallows for Condition 2 was demonstrated by a significant interaction between 'Boluses and Conditions,' significant at the 0.001 level ($F_{2,20} = 10.31$). Interactions between 'Subtests and Boluses' ($F_{10,20} = 1.85$), and 'Subtests and Conditions' ($F_{20,20} = 1.66$) were not significant.

The scores obtained by raters who considered themselves to be experienced were compared using Spearman's Rho and converted to Fisher's Z_r ($n = 2$, $r = 0.757$, $Z_r = 1$). A comparison of scores obtained by raters with minimal experience ($n = 2$) was also made using the same process ($r = .58$, $Z_r = 0.74$). The scores of raters who reported having some experience ($n = 5$) were compared using Spearman's Rho in a correlation matrix. The Rho scores were then converted to Pearson's Z_r scores and the mean Z_r was calculated (mean $Z_r = 0.90$). Though these figures suggest a trend towards better levels of agreement with increased experience, larger numbers of raters and better definitions of levels of experience are needed before meaningful conclusions can be drawn.

Discussion

As expected, the highest levels of agreement were attained under Condition 2, when raters were able to discuss their decisions, thus creating a degree of consensus before rating. The lowest levels of agreement occurred under Condition 1 where the raters assessed the first swallow in isolation, relying only on their own interpretation of the written Scale. In Condition 3, where the rating was performed independently after the group discussion of Condition 2, an expected weakening of the agreement reached for Condition 2 occurred. There was an apparent improvement in the amount of agreement in Condition 3 compared with Condition 1, for both liquid and semisolid bolus presentations, the differences were not statistically significant.

The findings of this study, combined with the comments collected during the discussions in Condition 2, provide valuable insights into the nature of the task and suggested the following modifications and clarifications to improve interrater reliability during the interpretation of videofluoroscopy.

The Timing of Observations

One reported source of variation between raters was the timing of observations. The assessment of which functions are largely one of the occurrences during a normal swallow, such as velar and hyoid elevation and elicitation of the swallow reflex, becomes more complicated when several swallows are required to clear a bolus. This situation creates numerous possible points for a judgment to be made during each bolus presentation. A common example of this situation is when patients with poor tongue function deliver the bolus into the pharynx in a piecemeal manner.

Confusion also exists as to the timing of judgments of pooling in the valleculae, pyriform sinuses, and at cricopharyngeus sites. Should these be rated before the reflex is initiated, between reflexes, or after the bolus has been swallowed? The raters agree that, in cases of multiple swallows per bolus presentation, the point of assessment should be clearly stipulated.

Bolus Consistency

Dantas et al. [8] noted that boluses of higher viscosity had slower flow rates than thinner boluses. Therefore, the slower transit time for semisolid boluses would enable the raters to have more time to observe the swallow, and this may account for the better levels of agreement that were reached when rating semisolid swallows in this study. There was general agreement among raters that slow motion replay should be used during interpretation, especially for spatial judgments such as site of reflex initiation and velar elevation.

Quality of Image

The clarity of the videofluoroscopic image may also influence the concordance of raters' assessments. When the X-rays pass through areas of low density, the brightness of the image increases. Flaring, or hyper-illumination, occur at body margins and distort the image, especially around the lips. Likewise, shading of the image can obscure the view in regions of higher density. The soft palate and cricopharyngeus are vulnerable to shading [9]. Inconsistencies in rater agreement of these aspects may be due in part to this problem.

At times, raters reported uncertainties when interpreting the videofluoroscopic image, i.e., the raters

found the X-ray images of the pharynx, pyriform sinuses, and cricopharyngeus difficult to differentiate. Similar difficulties were also apparent in the studies by Wilcox et al. [6] and Gibson and Phyland [5]. The retention of barium mixture further obscured the view of these structures. Easier recognition of the structures might have been achieved through the use of photographic and diagrammatic examples.

Rater's Experience

Previous relevant experience is known to influence clinical judgments [10] and this may have influenced their interpretation of videofluoroscopic images of swallowing. Experienced raters' awareness of the range of possible swallowing behaviors enables them to better recognize abnormalities of function. At present there are large variations in the training of students and clinicians in the interpretation of videofluoroscopy. Further investigation of this factor is required, with a need to define levels of experience and determine the affect of training and experience on rating.

Task Complexity

Judgments of complex functions are susceptible to bias [11]. The task of assessing swallowing using videofluoroscopy entails the possibility of multiple judgments throughout the swallowing process. Lessening task complexity by reducing the number of possible observations at each level of the Scale, (i.e., by retaining only those considered most salient) should improve interrater reliability. The process of simplifying the procedure requires a systematic series of validity studies.

Conclusions

This study has demonstrated the complexity of judgments involved in videofluoroscopic assessment of swallowing. Although widely used and currently the most effective means of assessing dysphagia, clinicians need to be mindful of the limitations with respect to reliability and therefore the validity of this method.

The development of a procedure for videofluoroscopic assessment of swallowing with demonstrable good interrater reliability should offer a much needed and more valid clinical tool. This study represents an important step towards this goal.

Acknowledgments. Thanks to the speech pathologists who participated in this study and the staff and patients at Bethlehem Hospital in Melbourne for their assistance. This work has been supported by The Motor Neurone Association of Victoria and The Bethlehem Griffiths Foundation.

References

1. Bishop WS: Videofluoroscopy of dysphagia patients. *Radiography Today 55:*15–17, 1989
2. Logemann JA: *Manual for the Videofluoroscopic Study of Swallowing.* London: Taylor and Francis, 1986
3. Price GJ, Jones CJ, Charlton RA, Allen CMC: A combined approach to the assessment of neurological dysphagia. *Clin Otolaryngol 12:*197–201, 1987
4. Ekberg O, Nylander G, Frans-Thomas F, Sjoberg S, Birch-Jensen M, Hillarp B: Interobserver variability in cineradiographic assessment of pharyngeal function during swallow. *Dysphagia 3:*46–48, 1988
5. Gibson E, Phyland D: Rater reliability of the modified barium swallow. *Aust J Comm Dis 23:*54–60, 1995
6. Wilcox F, Liss JM, Siegal GM: Interjudge agreement in videofluoroscopy studies of swallowing. *J Speech Hear Res 39:*144–152, 1996
7. Tukey JW: Comparing individual means in the analysis of variance. *Biometrics* (June), 99–114, 1949
8. Dantas RO, Dodds WJ, Massey BT, Kern MK: The effect of high- vs low-density barium preparations on quantitative features of swallowing. *Am J Radiol 153:*1191–1195, 1989
9. Beck TJ, Gayler BW: Image quality and radiation levels in videofluoroscopy for swallowing studies: a review. *Dysphagia 5:*118–128, 1990
10. Thomas S, Wearing A, Bennett M: *Clinical Decision-Making for Nurses and Health Professionals.* Marrickville, Australia, Harcourt Brace Jovanovich Group, 1991
11. Guildford JP: *Psychometric Methods.* Bombay, New Delhi: Tata McGraw Hill, 1978, p 304
12. Enderby P: Frenchay dysarthria assessment. *Br J Disord Commun 15*:165–173, 1982