Dysphagia 11:93–98 (1996)

Dysphagia

A Penetration-Aspiration Scale

John C. Rosenbek, PhD,[1] Jo Anne Robbins, PhD,[2] Ellen B. Roecker, PhD,[3] Jame L. Coyle, MA,[2] and Jennifer L. Wood, MS[2]

Departments of [1]Neurology, [2]Medicine, and [3]Biostatistics, William S. Middleton Memorial Veterans Hospital, University of Wisconsin School of Medicine, Madison, Wisconsin, USA

Abstract. The development and use of an 8-point, equal-appearing interval scale to describe penetration and aspiration events are described. Scores are determined primarily by the depth to which material passes in the airway and by whether or not material entering the airway is expelled. Intra- and interjudge reliability have been established. Clinical and scientific uses of the scale are discussed.

Key words: Dysphagia — Aspiration — Penetration — Scaling — Deglutition — Deglutition disorders.

Impaired swallowing results from abnormal changes in the structures or movements necessary for normal swallowing. Signs of dysphagia may vary from person to person and over time in the same person. They also vary in their clinical significance. A sign that attracts considerable clinical attention when it occurs is aspiration, defined generically as the passage of foreign material into the airway. Prevailing clinical wisdom has it that aspiration is clinically important for a variety of reasons, including the effect it may have on a person's health. It turns out, however, that the effects of aspiration on health are difficult to predict. Indeed, the one verity to emerge from several years of clinical practice and research is that aspiration does not affect all dysphagic persons equally [1]. The person's pulmonary, oral, and general health, mobility, cognition, frequency of aspiration, and type of material aspirated all can influence aspiration's effects [2]. Other influences may be at work as well, including the amount of aspiration, how far into the airway material passes, and whether or not the person is able to expel it.

Heretofore, clinicians and clinical researchers interested in investigating these and other influences have relied primarily upon notational methods of describing aspiration. These included recording whether or not aspiration has occurred and whether the occurrence is before, during, or after the swallow [3]. Some clinical scientists also estimate the percentage of the total bolus aspirated [4]. Such notations may have problems of reliability and validity, but they have been sufficiently robust to support significant recent advances in knowledge about aspiration's effects [5]. Further advances in understanding the effects of aspiration will depend, in part, on additional tools.

The 8-point Penetration-Aspiration Scale described here is offered as one such tool. The primary purposes of this paper are to (1) define critical terms, (2) describe the scale and its development, (3) report reliability data, and (4) describe selected present and future uses of the scale.

Definitions

Penetration is defined here as passage of material into the larynx that does not pass below the vocal folds. The amount of material, the depth of penetration, and whether all or a portion is subsequently expelled are potentially critical variables and deserve study, but are not part of the definition. *Aspiration* is defined as passage of material below the level of the vocal folds. Again, the amount, the distance the material passes into the trachea, and whether all or a portion is expelled are not part of this definition despite their potential clinical significance.

This work was performed at the William S. Middleton Memorial Veterans Hospital and the University of Wisconsin Clinical Science Center. This is publication number 94-10 of the Madison Geriatrics Research Education and Clinical Center

Offprint requests to: John C. Rosenbek, Chief, Audiology and Speech Pathology, William S. Middleton Memorial Veterans Hospital, 2500 Overlook Terrace, Madison, WI 53705, USA

Table 1. The original 9-Point Penetration-Aspiration Scale

1. Material does not enter the airway
2. Material enters the airway, remains above the vocal folds, and is ejected from the airway
3. Material enters the airway, remains above the vocal folds, and is not ejected from the airway
4. Material enters the airway, contacts the vocal folds, and is ejected from the airway
5. Material enters the airway, contacts the vocal folds, and is not ejected from the airway
6. Material enters the airway, passes below the vocal folds, and is ejected from the airway
7. Material enters the airway, passes below the vocal folds, and is ejected from the trachea into the larynx
8. Material enters the airway, passes below the vocal folds, and is not ejected from the trachea despite effort
9. Material enters the airway, passes below the vocal folds, and no spontaneous effort is made to eject the material

Table 2. Final version of the 8-Point Penetration-Aspiration Scale

1. Material does not enter the airway
2. Material enters the airway, remains above the vocal folds, and is ejected from the airway
3. Material enters the airway, remains above the vocal folds, and is not ejected from the airway
4. Material enters the airway, contacts the vocal folds, and is ejected from the airway
5. Material enters the airway, contacts the vocal folds, and is not ejected from the airway
6. Material enters the airway, passes below the vocal folds and is ejected into the larynx or out of the airway
7. Material enters the airway, passes below the vocal folds, and is not ejected from the trachea despite effort
8. Material enters the airway, passes below the vocal folds, and no effort is made to eject

Development of the Scale

Development of the scale began when 4 clinical scientists in the Veterans Administration/University of Wisconsin Swallowing Laboratory produced a ranked scale of goodness with regard to penetration and aspiration events or behaviors. The result was a 9-point scale shown in Table 1.

To determine if these 9 points described all the penetration and aspiration events in the videofluoroscopic swallowing examinations of a sample of dysphagic subjects, 4 judges used the scale to evaluate 75, 3-ml-thin, liquid bolus swallows, 5 from each of 15 subjects who were dysphagic because of multiple strokes. These swallowing evaluations were standardized so that fluoroscopy was initiated before the swallow and continued until a swallow was completed, as operationally defined by hyoid return to rest. Judges independently evaluated all swallows and were free to replay each swallow in any manner (frame-by-frame, continuous) and as frequently as necessary for confident judgment.

Two results emerged. First, the scale described all the penetration and aspiration events produced by the sample of 15 subjects. Second, one of the scores, number 5, never occurred. The experiment was replicated using three of the four original judges and two other previously assembled groups of research subjects: 40 normal, older adults and 12 head and neck cancer patients. For the normal subjects, three thin liquid bolus swallows were evaluated. For the head and neck cancer subjects, eight boluses of various types were evaluated. These groups were selected primarily because a future research goal is to investigate the Penetration-Aspiration scale's sensitivity to different patterns of swallowing abnormality that may reflect different underlying etiologies of dysphagia.

The results were the same. The scale described all the penetration and aspiration events displayed by these two groups, and one score, number 5, never occurred. In other words, none of these subjects ejected aspirated material completely out of the airway. In addition, a score of 4, indicating that material contacts the vocal folds and is then expelled from the larynx, occurred infrequently.

It was decided to reduce the 9-point scale to an 8-point scale by combining the descriptions of a former score of 5 with a former 7. Despite the relative rarity of number 4, it was retained because of its clinical importance. The final, 8-point version of the Penetration-Aspiration Scale appears in Table 2.

Description of the Scale

This 8-point version of the scale is multidimensional, meaning that more than one type of behavior is being judged. Depth of bolus invasion into the airway is a major dimension. Material (1) does not enter the airway, (2) enters the larynx but stays above the vocal folds, (3) enters the larynx to the level of the vocal folds, or (4) passes below the vocal folds. The swallower's response to the bolus is a second dimension. Material is completely expelled, partially expelled, or not expelled. The scale is represented schematically in Figure 1.

The scale was also assembled to be ordinal. Each behavior identified by scores 2 through 8 is assumed to be a more severe sign of dysphagia than the behavior identified in the preceding score. Aspiration is judged to be more severe than penetration. Therefore, aspiration is scored 6, 7, or 8. Penetration, on the other hand, is scored either 2 or 3 if residue remains above the vocal folds and 4 or 5 if residue courses to the level of the vocal folds. Whether or not material is ejected from the airway also

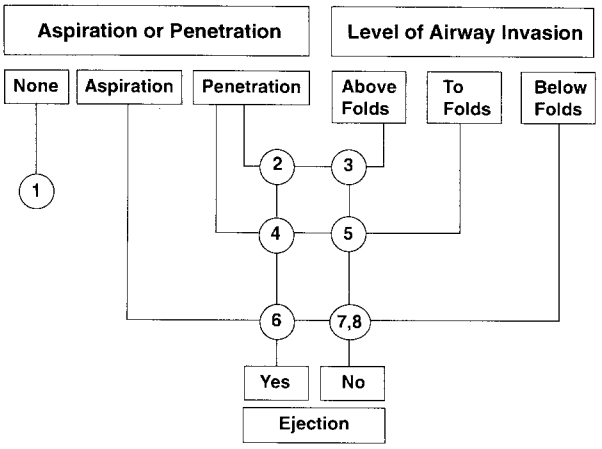

Fig. 1. Schematic representation of the Penetration-Aspiration Scale.

contributes to a judgment about severity. The influence of this dimension can be seen by comparing values 4 and 5. It is judged that material penetrating to the level of the vocal folds and not ejected (score of 5) creates a more serious condition than if material passing to that level is ejected (score of 4). This dimension was judged a less important influence on severity than is the penetration-aspiration dimension, however. For example, even if aspirated material is expelled partially or totally from the airway (score of 6), that condition is judged to be more serious than if material penetrates and is not expelled (score of 5). A third influence on the scale's ordinality, but only at the most severe end of the scale, is whether or not a person responds to aspiration. The most severe condition is aspiration without either a reflexive or conscious attempt to expel it (score of 8), a condition identified clinically as silent aspiration.

Reliability of the Scale

To measure reliability, the videofluoroscopic images of 75 swallows elicited as part of a standardized experimental protocol [6] were copied in random sequence onto a videotape. The 75 swallows were elicited from the original 15 dysphagic, stroke subjects. Four experienced judges who had previously reviewed the definitions of each of the 8-scale scores viewed and assigned a score to each of the 75 swallows to assess interjudge reliability. Two weeks later, to assess intrajudge reliability, the 4 judges again reviewed the tape and assigned scores to each of the 75 swallows. Judges were allowed to view each swallow at a variety of speeds and as often as necessary for confident judgment.

Table 3 shows two kinds of data. The overall frequency of scale scores in this sample of stroke subjects

Table 3. Cross-classification of scores given by 4 judges when 75 swallows were graded a second time by the same judge

		Second grading score									
		1	2	3	4	5	6	7	8	Total	%
First grading score	1	**71**	11	7						89	30
	2	2	**34**	16		1				58	19
	3	1	5	**44**		8	2			60	20
	4				**0**	1				1	.3
	5	1		1	2	**9**	3		3	19	6
	6						**3**	1	2	6	2
	7	1						**34**	3	38	13
	8							1	**28**	29	10
	Total	81	50	68	2	19	8	36	36	300	100
	%	27	16	23	1	6	3	12	12		

for both gradings by the 4 judges appear along the right and bottom margins of the table. The most frequently assigned scores were 1–3 followed by 7 and 8. Very few gradings received a score of 4 (3/600 = 0.5%).

Table 3 also provides the pattern of agreement between the first and second grading of each swallow by the same judge. The 4 judges assigned the same score to the swallow when they graded it the second time in 223 of 300 replicate gradings (74% agreement). When an identical score was not given on the second grading, a score within 1 unit (50/300 = 17%) or 2 units (19/300 = 6%) were the most frequent. Scores differing by more than 2 units were given in only 8 of the 300 replicate gradings (3%) by the 4 judges combined. Among the 77 replicate gradings which did not agree, scores on the second grading were higher than those for the first grading in 58 (75%), suggesting a tendency of the judges to alter their scoring behavior on the second grading.

The number of times the two gradings by a judge were in agreement is shown for each judge in Table 4. These range from 47 of 75 gradings (63%) for Judge 1 to 63 of 75 gradings (84%) for Judge 3. In order to assess whether some of the scale scores seemed to be more difficult to grade than others, score-specific, intraclass kappa coefficients (or $k_{(i)}$ using the terminology of Bloch and Kraemer [7]) were computed to measure intrajudge agreement on individual scores for each judge and over all judges. These values are also shown in Table 4. Scores of 7, 8, and 1 were the most reliably graded, as indicated by kappas which exceed 0.75. Other scores were less reliably graded, and a score of 4 was too infrequently given for reliability to be adequately assessed.

The agreement between judges was descriptively assessed in a similar fashion. The 4 judges comprise 6 judge pairs, whose agreement in scoring on the first grading by each judge is summarized in Table 5. The interjudge agreement is comparable to the intrajudge agreement. When the scores given by the 2 judges were not the same, they usually only differed by 1 or 2 units

Table 4. Intrajudge agreement by judge on two gradings of 75 swallows

	Agreement		Intraclass Kappa (κ_I) by scale score							
Judge	n	%	1	2	3	4	5	6	7	8
1	47	63	0.60	0.47	0.41	—	0.46	—	0.80	0.75
2	57	76	0.65	0.47	0.74	—	0.55	—	0.93	0.93
3	63	84	0.87	0.77	0.76	—	0.79	0.38	0.88	0.78
4	56	75	0.88	0.46	0.50	—	0.21	0.47	1.0	0.88
Overall	223	74	0.77	0.55	0.60	—	0.44	0.42	0.91	0.84

Table 5. Interjudge agreement

Judge pair	1–2	1–3	1–4	2–3	2–4	3–4
Two scores agree						
n	45	54	43	56	56	56
% (of 75)	60	72	57	75	75	75
Number of scores that differ by						
1	23	16	16	17	10	9
2	6	4	10	2	2	6
3	1	1	4	0	7	2
>3	0	0	2	0	0	2

Table 6. Interjudge intraclass Kappa coefficients (κ_I) by scale score

Judge	Scale score							
Pair	1	2	3	4	5	6	7	8
1–2	0.45	0.24	0.49	—	0.57	—	0.80	0.84
1–3	0.62	0.54	0.74	—	0.47	—	0.82	0.82
1–4	0.52	0.42	0.25	—	0.24	—	0.87	0.78
2–3	0.79	0.47	0.67	—	0.31	—	0.88	0.84
2–4	0.94	0.50	0.57	—	0.12	—	0.94	0.80
3–4	0.85	0.60	0.41	—	0.41	—	0.94	0.78
Average:	0.70	0.46	0.52	—	0.35	—	0.88	0.81

on the scale. Intraclass kappa coefficients were computed for each scale score and judge pair and these are shown in Table 6. On average, the interjudge values are somewhat lower than the intrajudge values, but again suggest that scores of 7, 8, and 1 are more reliably assessed than scores in the middle of the scale.

The intraclass kappa coefficient used here is closely related to Cohen's kappa [8] but better reflects agreement as opposed to association [7]. In computing kappa, grading disagreements are all treated as equally bad. An index of reliability which reflects the fact that the Penetration-Aspiration Scale categories are ordered, and considers large score disagreements more serious than small, would be preferable. Also, the swallowing assessment of a subject usually comprises multiple swallows, with a mean over swallows or other summary assessment made for the subject. It therefore becomes more relevant to assess reliability of a subject assessment, based

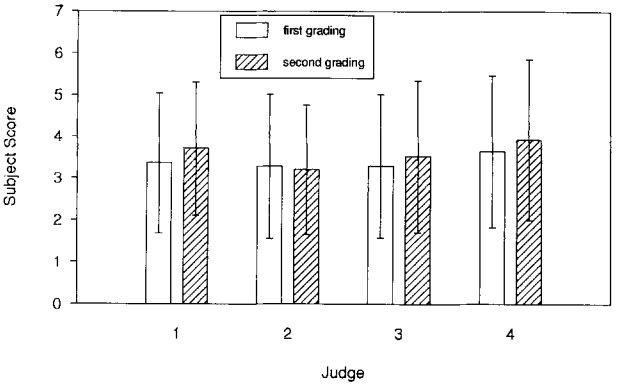

Fig. 2. Mean (± standard deviation) subject scores by judge and judge replicate for the 15 subjects. Subject scores are average scores for 5 swallows. There is a significant judge by judge replicate interaction ($p < 0.05$) by repeated measures ANOVA. Judges 1, 3 and 4 gave significantly higher scores on the second grading. The differences among judges were also significant ($p < 0.05$) for the first and second replicate.

on a subject score that is the average scale score for five swallows. With these considerations in mind, an intraclass correlation coefficient based on a two-way random, effects analysis of variance (ANOVA) model (specifically, the ICC (2,1) defined by Shrout and Fleiss [9]) was chosen as the summary index of intra- and interjudge reliability across all categories of the scale for subject evaluations made using mean scores.

Figure 2 shows the average mean scores for the 15 subjects by judge and judge replicate and the variation (standard deviation) in mean score observed among the subjects. Three of the 4 judges appear to have given somewhat higher scores on the second grading. Repeated measures ANOVA was used to assess the statistical significance of any differences observed among judges and between gradings by each judge, using Huynh-Feldt adjusted p values [10]. There was a statistically significant ($p = 0.02$) judge by judge replicate interaction, indicating that the difference between the first and second replicate varies according to the judge. Judges 1, 3, and 4 each gave significantly higher (more severe) scores on the second grading than the first, whereas Judge 2 scores did not differ between the first and second grading. There

were also significant differences among judges for both the first ($p = 0.03$) and second ($p = 0.002$) replicate, with Judge 4 scores tending to be systematically higher than those of the other judges on both replicates. Judge 1 scores were significantly greater than Judge 2 scores on the second replicate.

Despite this evidence of some systematic variation among judges, and over time for individual judges, the reliability of the scale is very high relative to the variation observed between subjects. The interjudge intraclass correlation coefficient calculated based on the first replicate of each judge is 0.96 with an approximate 95% confidence interval, calculated according to the formula given by Shrout and Fleiss [9] of 0.91–0.98. The intraclass correlation coefficient, which varies between 0 and 1 with 1 indicating high reliability, can be interpreted as the proportion of the variability in the measurement which is due to true differences between subjects as opposed to variability among judges. It can also be interpreted as the correlation between two measures on the same subject by randomly selected judges. Intrajudge intraclass correlation coefficients for each judge ranged from 0.95 for Judge 1–0.97 for Judge 3, which can be interpreted as the correlation between two measures on the same subject by the same judge. In other words, judges are almost as consistent with each other as they are with themselves.

Discussion

The Penetration-Aspiration Scale was developed to provide reliable quantification of selected penetration and aspiration events observed during videofluoroscopic swallowing evaluations. It does not quantify all such events nor was it intended to. Users are left to use other systems to specify the amount and timing of penetration and aspiration events. Nor can the scale substitute for other perceptual measures of swallowing tested videofluoroscopically. Depending on the examiner's purposes, duration measures, notation about pooling and coating, piecemeal deglutition, abnormal movements, and other traditional signs of abnormal swallowing will remain critical. The scale is offered as one tool to be included as part of a total swallowing assessment battery.

Trained, reliable clinicians can use the scale to clinical advantage. The training itself may make them better observers of videofluoroscopic images. Communication among similarly trained clinicians can be more efficient, because the path of the bolus on one or a series of swallows can be summarized by one or more numbers. Such standardization may have the effect of allowing clinicians to compare shared patients or the characteristics of different practices. The scale can help make reports of penetration and aspiration events more precise. Two signs—penetration and aspiration—have become eight signs. To use the scale reliably, the clinician must note how far into the airway material passes and whether or not it is expelled. Simply noting penetration or aspiration is not enough. Finally, trained, reliable clinicians can also use the scale to improve the training of students and new clinicians.

The scale would appear to offer some advantages to the clinical researcher as well. One of the authors (JR) is testing the hypothesis that the scale will aid the differentiation of normal and abnormal swallowers and may even help in the differential diagnosis of some swallowing-impaired populations such as those with dementia. Preliminary data confirm that normal older swallowers sometimes earn scores of 2 and 3. Using the scale, along with other measures, of course, may also help researchers discover why some patients who aspirate get sick and some do not. The scale is also a potentially powerful outcome measure for clinical trials designed to investigate the efficacy of various swallowing treatments. Unlike at least some duration measures and perhaps even some of the other traditional signs of dysphagia, penetration and aspiration have a clear clinical significance. Most researchers would agree that a reduction in the number of times a patient penetrates or aspirates is a sign of improvement. The scale may be equally useful to the clinician interested in demonstrating a functional change in the individual patient.

Research uses of the scale raise critical issues about the scale's characteristics. An experiment is planned to evaluate the scale's ordinality using 25 independent judges experienced in swallowing evaluation and assessment but who had no hand in the scale's development. Another experiment will examine the degree to which the present scale can be considered an interval scale in which all distances between adjacent scores are equal. This can be established by developing an interval scale using the 8 categories in a paired comparisons paradigm following experimental and statistical procedures developed by Guilford [11]. Another experiment will determine the relationship between scores of videofluoroscopic examinations for a group of dysphagic patients obtained using the original and the experimentally derived interval scale. These experiments will affect the confidence with which researchers use summary statistics of performance on this scale to answer experimental questions about such things as treatment efficacy.

Summary

Accurate diagnosis of medical conditions often requires procedures that increase the clinician's sensitivity to aspects of a disorder. Widespread implementation of such procedures may offer a variety of related advantages including improved communication among professionals

as well as providing a reference by which treatment effectiveness for the targeted disorder may be assessed. The Penetration-Aspiration Scale is a newly developed tool for such purposes. Acceptable intra- and interjudge reliability for the scale have been established sufficient enough to support its introduction into clinical practice. The advantages of its incorporation into clinical practice as well as research paradigms are numerous, but information gained through continued evaluation of scale construction may increase its utility even beyond the currently apparent applications.

Acknowledgment. This work was supported by the Department of Veterans Affairs Rehabilitation, Research and Development grant E728-GA.

References

1. Kirsch CM, Sanders A: Aspiration pneumonia: medical management. *Otolaryngol Clin North Am 21:*677–689, 1988
2. Langmore SE: Managing the complications of aspiration in dysphagic adults. *Semin Speech Language 12:*199–207, 1991
3. Logemann JA: *Evaluation and Treatment of Swallowing Disorders.:* San Diego: College-Hill Press, 1983
4. Rademaker AW, Pauloski BR, Logemann JA, Shanahan TK: Oropharyngeal swallow efficiency as a representative measure of swallowing function. *J Speech Hear Res 37:*314–325, 1994
5. Martin BJW, Corlew MM, Wood H, Olson D, Golopol LA, Wingo M, Kirmani N: The association of swallowing dysfunction and aspiration pneumonia. *Dysphagia 9:*1–6, 1994
6. Lof GL, Robbins J: Test-retest variability in normal swallowing. *Dysphagia 4:*236–242, 1990
7. Bloch DA, Kraemer HC: 2 × 2 Kappa coefficients: measures of agreement or association. *Biometrics 45:*269–287, 1989
8. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas 20:*37–46, 1960
9. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. *Psychol Bull 86:*420–428, 1979
10. SAS Institute Inc.: *SAS/STAT User's Guide:* vol 2 GLM-VARCOMP, version 6, 4th ed. Cary, NC: SAS Institute Inc, 1990
11. Guilford JP: *Psychometric Methods,* 2nd ed. New York: McGraw-Hill, 1954