

OSHA 2007: Do You See What I See: Results of the Reliability Survey

Description:

Clinicians with varying levels of experience were asked to rate 10 swallowing clips of single swallows. Clips were played twice, once in real-time and once in slow-motion, and broadcast on a screen in a semi-darkened auditorium. Clips included normal and abnormal swallows of various consistencies and with differing levels of magnification. No background information about each clip was provided. Individuals were asked to rate the clip using a binary system of normal or abnormal for each of 10 parameters. No additional guidance was provided on the first scoring trial. Subsequently each of the parameters was discussed during a presentation and the audience was asked to discuss their criteria for scoring. The 10 clips were played again at the end of the presentation and the audience was asked to re-rate the clips using the same scoring scheme.

A total of 41 scorers handed in rating forms. Of these, 2 had scoring data from the first but not the second trial and were excluded. Thus, there were a total of 39 scorers rating 10 parameters of 10 swallows across 2 conditions = a total of 7,800 decisions.

Background characteristics of the audience:

Individuals were asked to provide information about their background, experience and current employment situation, as follows.

- Current status: SLP (79%), student (18%), other (3%).
- Number of years practicing as SLP: none (18%), 1-5 (21%), 6-10 (15%), 11+ (46%).
- Number of years of dysphagia experience: none (21%), 1-5 (23%), 6-10 (15%), 11+ (41%).
- Training in dysphagia (list all): workshops (74%), graduate class/seminar (62%), videos/CD's (46%), practicum (44%), fellowship (18%), none (5%).
- Primary employment setting (choose one): hospital (31%), SNF (15%), outpatient clinic (15%), none (13%), school (10%), rehab facility (8%), home health (5%), other (3%).
- All employment settings (list all): hospital (44%), outpatient clinic (33%), SNF (23%), rehab facility (18%), home health (15%), school (13%), rehab facility (9%), university (10%), private practice (8%), other (5%).
- Primary caseload: adult (53%), adult & pediatric (24%), pediatric (13%), none (11%).
- Number of MBS studies performed per week (on average): none (47%), <1 (17%), 1-5 (28%), 5+ (8%).

Statistics summary:

Scores were summarized as either 1 (present), 2 (absent), or 3 (no decision). Generalized kappa was then calculated in SPSS using the macro available online at:

<ftp://ftp.spss.com/pub/spss/statistics/nichols/macros/mkappasc.sps>

Guidelines for kappa:

The kappa value ranges from 0 (no agreement) to 1 (complete agreement). The values are often interpreted as follows.

Kappa value	Level of agreement
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very Good

Prior to discussion:

	kappa	significance
Reduced lingual control	.38	<.001
Oral residue	.38	<.001
Vallecular residue	.63	<.001
Pyriiform residue	.53	<.001
Hypopharyngeal residue	.54	<.001
Epiglottis dysfunction	.41	<.001
Penetration	.57	<.001
Aspiration	.48	<.001
Reduced hyolaryngeal excursion	.31	<.001
Cricopharyngeal prominence	.09	NS

After discussion:

	kappa	significance
Reduced lingual control	.43	<.001
Oral residue	.43	<.001
Vallecular residue	.62	<.001
Pyriiform residue	.62	<.001
Hypopharyngeal residue	.54	<.001
Epiglottis dysfunction	.46	<.001
Penetration	.58	<.001
Aspiration	.69	<.001
Reduced hyolaryngeal excursion	.32	<.001
Cricopharyngeal prominence	.09	NS

Results:

In the first trial only one parameter was rated in the “good” range, namely “vallecular residue” ($\kappa=.63$). Five were in the moderate range (“pyriiform residue,” “hypopharyngeal residue,” “epiglottis dysfunction,” “penetration,” and “aspiration”) and 3 in the “fair” range (“reduced lingual control,” “oral residue,” and “reduced hyolaryngeal excursion”). One parameter, “cricopharyngeal prominence,” demonstrated essentially no agreement.

In the second trial 3 parameters were rated in the “good” range, namely “vallecular residue” ($\kappa=.62$), “pyriiform residue” ($\kappa=.62$), and “aspiration” ($\kappa=.69$). There were 5 parameters in the “moderate range” (“reduced lingual control,” “oral residue,” “hypopharyngeal residue,” “epiglottis dysfunction,” and “penetration”) and only one parameter in the “fair” range (“reduced hyolaryngeal excursion”). “Cricopharyngeal prominence” continued to demonstrate no agreement.

Discussion:

Previous studies of reliability of interpretation of the modified barium swallow study have shown low levels of agreement for radiologists, speech pathologists, and for swallowing centers (Ekberg, et al., 1988; Kuhlemeier, Yates, & Palmer, 1998; McCullough, et al., 2001; Stoeckli, Thierry, Seifert, & Martin-Harris, 2003; Wilcox, Liss, & Siegel, 1996). In one study of 3 expert clinicians, the only parameter rated reliably was the presence of penetration/aspiration (McCullough, et al., 2001). Interrater reliability has not been shown to improve even with the use of measurement software (Dyer, Leslie, & Drinnan, 2007). Reliability has been shown to improve, however, with group discussion (Scott, Perry, & Bench, 1998). Although it is known that high levels of agreement for swallowing judgements are routinely reported at research facilities, there is limited research about the relationship between the accuracy of observations and the number of hours of training received (Logemann, et al., 2005).

In this study, speech pathologists and speech pathology students with a wide range of experience were asked to rate clips of single swallows from modified barium swallow studies. Initially the participants were asked to rate 10 parameters of swallowing as either normal or abnormal, using a scoring system used in previous research (Perlman, Booth, & Grayhack, 1994; McCullough, et al., 2001). Subsequently research about each of the parameters was presented and clinicians in the audience discussed their own standards for rating each of these findings. No standardized definitions or scoring guidelines were established before the audience re-rated the same 10 clips.

There are a number of characteristics of this task which may have contributed to poor agreement: there were only two opportunities to view each clip before scoring, no background information about the subject or the bolus consistency was provided, image quality and magnification were not uniform resulting in differences in the ability to view the swallowing anatomy & physiology, and there were several parameters to score. In addition, there was minimal information provided about the parameters prior to the first scoring trial. The audience was also made up of individuals with a wide range of experience. All of these may have contributed to the fact that only one parameter was scored in the “good” range on the first trial.

In the second trial there was general improvement with five parameters increasing their reliability rating somewhat, either from “fair” to “moderate” or “moderate” to “good.” Encouragingly, there was much stronger agreement for the parameter of “aspiration” on the second trial. Two parameters that continued to demonstrate low levels of agreement were “hyolaryngeal elevation” and “cricopharyngeal prominence.” It was notable that the 3 parameters rated in the “good” range on the second trial were all bolus-related rather than physiologic parameters (i.e. “vallecular residue,” “pyriform residue,” and “aspiration”). This may be related to the relative speed of the movements of swallowing.

There are a number of explanations for the relative improvement seen between the first and second trials: repeated viewings of the same clips, increased familiarity with the scoring scale, and discussion of the parameters themselves with other examples of normal and abnormal findings during the seminar. Our findings mirror those of previous research which has shown discussion may improve inter-rater reliability in modified barium swallow study interpretation (Scott, Perry, & Bench, 1998).

Feedback from our participants included the observations that the 10 parameters were too numerous, some of the terms used were unfamiliar, and that orientation to the anatomy prior to the first trial or additional viewings of each clip would have been helpful. While some reported that re-rating the clips on the second trial was easier, others reported

that it was more difficult as they were now aware that the scoring criteria of other members of the audience were different from their own.

Conclusion:

Discussion of ten swallowing parameters improved reliability of ratings in an audience with a vast range of experience, from no dysphagia experience or training at all to many years of dysphagia practice. This occurred in the context of a relatively-challenging scoring task. These data suggest that this approach may be an appropriate way to begin the process of increasing reliability in the interpretation of the modified barium swallow.

References:

- Dyer, JC, Leslie, P, & Drinnan, MJ. 2007. Objective computer-based assessment of valleculae residue: Is it useful? *Dysphagia*, in press.
- Ekberg, O, Nylander, G, Fork, F-T, Sjoberg, S, Birch-Iensen, M, & Hillarp, B. 1988. Interobserver variability in cineradiographic assessment of pharyngeal function during swallow. *Dysphagia*, 3: 46-48.
- Kuhlemeier, KV, Yates, P, & Palmer, JB. 1998. Intra- and interrater variation in the evaluation of videofluorographic swallowing studies. *Dysphagia*, 13: 142-147.
- Logemann, JA, Williams, RB, Rademaker, A, Pauloski, BR, Lazarus, CL, & Cook, I. 2005. The relationship between observations and measures of oral and pharyngeal residue from videofluorography and scintigraphy. *Dysphagia*, 20: 226-231.
- McCullough, GH, Wertz, RT, Rosenbek, JC, Mills, RH, Webb, WG, & Ross, KB. 2001. Inter- and intrajudge reliability for videofluoroscopic swallowing evaluation measures. *Dysphagia*, 16: 110-118.
- Perlman, A.L., Booth, B.M., & Grayhack, J.P. 1994. Videofluoroscopic predictors of aspiration in patients with oropharyngeal dysphagia. *Dysphagia*, 9: 90-95.
- Scott, A, Perry, A, & Bench, J. 1998. A study of interrater reliability when using videofluoroscopy as an assessment of swallowing. *Dysphagia*, 13: 223-227.
- Stoeckli, SJ, Thierry, AGM, Seifert, B, & Martin-Harris, BJW. 2003. Inter-rater reliability of videofluoroscopic swallow evaluations. *Dysphagia*, 18: 53-57.
- Wilcox, F, Liss, JM, & Siegel, GM. 1996. Interjudge agreement in videofluoroscopic studies of swallowing. *Journal of Speech and Hearing Research*, 39: 144-152.