Dysphagia 16:110–118 (2001)
DOI: 10.1007/s004550010004

Dysphagia

Inter- and Intrajudge Reliability for Videofluoroscopic Swallowing Evaluation Measures

Gary H. McCullough, PhD,[1] Robert T. Wertz, PhD,[2,3] John C. Rosenbek, PhD,[4] Russell H. Mills, PhD,[2,5] Wanda G. Webb, PhD,[2] and Katherine B. Ross, PhD [6]

[1]Department of Audiology and Speech Pathology, University of Tennessee, Knoxville, Tennessee; [2]Vanderbilt University, Nashville, Tennessee; [3]VA Medical Center, Nashville, Tennessee; [4]University of Wisconsin; and VA Medical Center, Madison, Wisconsin; [5]VA Medical Center, Murfreesboro, Tennessee; and [6]VA Medical Center, Phoenix, Arizona, USA

Submitted September 1, 1999; accepted August 23, 2000 with revision

Abstract. Interjudge reliability for videofluoroscopic (VFS) swallowing evaluations has been investigated, and results have, for the most part, indicated that reliability is poor. While previous studies are well-designed investigations of interjudge reliability, few reports of intrajudge reliability are available for VFS measures derived from frame-by-frame analysis that clinicians typically employ. The purpose of this study was to examine the inter- and intrajudge reliability of VFS examination measures commonly used to assess swallowing functions. No training to criteria occurred. VFS examinations were conducted on 20 patients who had suffered a stroke within six weeks and had no structural abnormalities or tracheostomies. Three clinical judges served as subjects and rated the VFS examinations from videotape using frame-by-frame analysis. A clinician's repeated review of measures employed in the 20 examinations indicated high intrajudge reliability for a number of measures, suggesting that an experienced clinician may employ consistent standards for rating certain VFS measures across patients and time. These standards appear to vary among clinicians and yield unacceptable interjudge reliability. The need to train clinicians to criteria to improve interjudge reliability is discussed.

Key words: Dysphagia — Videofluoroscopy — Reliability — Adults — Deglutition — Deglutition disorders.

Interjudge reliability for measures employed in videofluoroscopic (VFS) swallowing evaluations were investigated [1–4]. Results of those investigations demonstrated poor interjudge agreement for most measures employed. One study [4] provided evidence that training to criterion can improve interjudge reliability. The authors examined interjudge reliability under varying levels of communication. When judges worked alone, reliability data were as poor as other investigations have reported. When judges discussed how to rate measures prior to the actual ratings, reliability improved. Similar evidence emerged from a study by Perlman et al. [5] that attempted to define the clinical correlates of dysphagia. The authors reported good interjudge reliability for the VFS measures employed. Judges in that investigation were pretrained to a criterion performance prior to the collection of reliability data. More recently, Smith et al. [6] also reported good interjudge reliability among pretrained judges for reporting the presence of aspiration from VFS studies. Other investigations targeting information related to swallowing function and aging [7], the effects of a sour bolus [8], and clinical indicators of dysphagia [9] have reported good interjudge reliability as well. Those investigations, however, did not report whether pretraining to criterion performance was used. Duration measures are not yet typically employed in clinical practice [10], but interjudge reliability has been investigated. For

Subjects for this project were recruited from VA Medical Centers in Nashville and Murfreesboro, TN, and Vanderbilt University Medical Center, Nashville, TN.

Correspondence to: Gary H. McCullough, Ph.D., University of Tennessee, Department of Audiology and Speech Pathology, 457 S. Stadium Hall, Knoxville, TN 37996, USA. E-mail: gmccullo@utk.edu

example, high interjudge reliability for a number of duration measures was established in two studies involving VFS data from normal patients [11,12]. Good interjudge reliability has also been reported for an 8-point penetration-aspiration scale [13]. Judges in those three investigations were pretrained to criterion performance. Thus, the results of interjudge reliability reports vary. Additionally, interjudge reliability on VFS measures appears to improve when judges are pretrained to criterion performance [4,6,11–13].

While there are a number of interjudge reliability investigations, only one [3], to our knowledge, examined intrajudge reliability for measures that clinicians typically employ in their evaluations [10]. Results of that investigation [3] demonstrated that intrajudge reliability was poor. However, the authors indicated that reliability was not obtained from VFS frame-by-frame analysis. Investigations of duration measures [11,12,14] and an 8-point penetration–aspiration scale [13] have reported good intrajudge reliability as well. However, all judges were pretrained to criterion performance. Thus, for the majority of VFS measures clinicians typically employ [10], few intrajudge reliability data are available on clinicians who were not pretrained to criterion performance.

The purpose of this study was to examine clinicians' inter- and intrajudge reliability on VFS examination procedures and measures commonly used in the assessment of the swallowing function [10]. None of the participants (clinical judges) were pretrained to criterion performance. Thus, results of this investigation provide insight into potential reliability problems for clinicians attempting to rate VFS measures based on definitions without pretraining. VFS frame-by-frame analysis was employed and quantitative data were derived.

Methods

The material for the reliability studies was provided by 20 patients who suffered a stroke and who received a VFS swallowing evaluation: 14 from the Veteran's Administration Medical Center (VAMC) in Nashville, Tennessee; 5 from Vanderbilt University Medical Center; and 1 from the VAMC in Murfreesboro, Tennessee. The 14 patients from VAMC, Nashville, were consecutive stroke patients. Six additional patients were referred from the other two hospitals by clinicians who identified them as individuals who met selection criteria. All patients were recruited over a six-month period. All had suffered a stroke within six weeks of the time of examination (16 were within two weeks post-onset). Patients with previous strokes were included as long as no swallowing problems were reported to exist from the prior stroke. Data were not included for any patients who (1) had an anatomical/structural deviation which would affect swallowing or (2) had a current or recent tracheostomy. Patient descriptive data are located in Table 1. The patients' mean age was 67.8 years and the mean number of days post-onset was 7. The sample comprised 17 males and 3 females. Locations of the patients' lesions varied throughout cortical and subcortical areas but were predominantly unilateral. No brainstem lesions occurred in this sample. Fourteen of the 20 patients had some penetration or aspiration on at least one swallow, and 11 of the 20 had penetration or aspiration on more than one swallow. Six of the 20 were recommended for either a change in diet or the use of a compensatory strategy to decrease aspiration. Thus, some patients in our data sample were dysphagic.

Table 1. Descriptive data for patients who received a videofluoroscopic swallowing evaluation

Participant	Age/Gender	DPO[a]	Stroke(s) location[b]
1.	64/M	2	L-Frontoparietal
2.	40/M	1	R-Thalamus
3.	62/M	21	Cerebellar hemorrhage
4.	75/M	16	R-MCA distribution
5.	69/M	2	L-Frontoparietal
6.	83/F	42	L-Hemisphere
7.	64/M	14	R-Occipital
8.	65/M	7	R-Hemisphere (unspecified)
9.	75/M	1	L-Frontal
10.	63/M	1	L-Parietal/occipital
11.	75/M	3	Questionable location; Hx bilateral strokes
12.	48/F	6	R-Frontal
13.	64/M	2	Questionable location; L-frontal and occipital
14.	54/M	2	R-Parietal, subcortex, and corona radiata
15.	70/M	2	R-White matter
16.	63/M	7	R-Frontoparietal hemorrhage
17.	96/F	2	L-MCA distribution
18.	81/M	4	R-Temporal/thalamus
19.	72/M	4	R-Frontal; previous R-frontoparietal
20.	73/M	1	L-Occipital extension of old L-MCA distribution

[a]DPO = Days post-onset.
[b]L = left, R = right, MCA = middle cerebral artery, Hx = history.

All procedures were approved by the Committees for the Protection of Human Participants at Vanderbilt University and in both VA Medical Centers. Informed consent was obtained from all patients prior to the initiation of any of the examination procedures.

Participants

Participants in this study were three speech-language pathologists (SLPs). Each held the certificate of clinical competence from the American Speech-Language-Hearing Association, and each had obtained at least 300 hours of experience in evaluating and managing swallowing disorders. Each clinician had obtained at least 200 hours of experience with VFS swallowing evaluation procedures, though some of the duration measures employed in this investigation were new to them. Participant 1 was the primary study clinician. Participants 2 and 3 were the other clinical judges.

Design

Videofluoroscopy Examination

The protocol for the VFS evaluations was developed from a survey of clinicians' preferences and practices in conducting VFS examinations [10]. The most commonly used measures, based on the survey data, were included in the reliability analysis.

The primary study clinician administered each VFS examination. Each patient was seated upright in a wheelchair or stretcher chair for the duration of the study. Patients in VAMC, Nashville, were examined with a mobile C-arm x-ray system (Model 9400, OEC-Diagnostics, Inc., Salt Lake City, UT) run by a radiology technologist. Each study was recorded with a Panasonic Super VHS AG-1960 Pro Line Multiplex videocassette recorder with an attached digital videotimer (Model VC436, TEL Video Products, Ann Arbas, Michigan). The one VAMC, Murfreesboro, patient was examined with a Phillips Super 80CP fluoro unit, and the study recorder was a Panasonic A66300 VCR. At Vanderbilt University Medical Center, a Siemens fluoro unit (Model 8842437G5275) with a 40-in. fixed tower was used with a Panasonic AG6300 MD videocassette recorder.

Unless the study was terminated according to predetermined "bailout" criteria, four consistencies were administrered to each patient: two 5-cc thin liquids with a viscosity of 14 cP (centipoise) (E-Z-HD barium sulfate powder for suspension and water at 50/50); two 10-cc thin liquids with the same viscosity; two 5-cc thick liquids (Welch's grape juice, Thicken-Up, and barium sulfate) with a viscosity of 187 cP; two 10-cc thick liquids with the same viscosity; two purees [Musselman's applesauce (4 oz) and barium sulfate (2 tbsp)]; and two solids (1/4 Lorna Doone cookie with E-Z Paste Esophageal Cream). If, for clinical management purposes, any compensatory maneuvers were attempted, they occurred after all of the above swallows were either completed or aborted. None of the videofluoroscopic information collected for management purposes was included in the study data.

VFS Videotape Review

Each participant (clinical judge)—one to determine intrajudge reliability and three to determine interjudge reliability—viewed videotaped VFS studies in collections of 3–5 at a time and independently recorded his/her ratings on a data sheet. Judges had no knowledge of the participants whose tapes were being reviewed. Measures derived from the survey study [10] and evaluated for reliability in this investigation are listed and described in Appendix A. Most measures were rated on a binary scale: normal–abnormal or present–absent. An 8-point penetration–aspiration scale [13] was used in addition to the binary rating of penetration–aspiration. Timed duration measures [in milliseconds (ms) defined in Appendix A] were derived by calculating the differences between specified starting and stopping points based on anatomical and physiologic markers [11,12,14]. If a measure could not be rated from a patient's videotape, the clinician circled "CNA" (Cannot Assess) on the response form.

Interjudge reliability was determined by comparing the responses made on the data sheets by all three participants (clinical judges). At least one week after the original viewing, the primary study clinician reanalyzed each of the VFS examinations and recorded all measurements on a new data sheet. Intrajudge reliability was determined by comparing his original ratings with his ratings from the second viewing.

Analysis of Data

All data were entered on spreadsheets and analyzed using SPSS 10.0 for Windows. The following analyses determined intrajudge reliability: for all binary ratings, Cohen's kappa; for all duration measures, Pearson's product moment correlations; and for the 8-point penetration–aspiration scale, Kendall's tau correlations. To determine interjudge reliability, the following analyses were performed: for all binary ratings, group kappas; and for all duration measures and the 8-point penetration–aspiration scale, intraclass correlation coefficients. The intraclass correlation coefficient (ICC) is based on a two-way random effects analysis of variance (ANOVA) model, as defined by Shrout and Fleiss [15]. A two-way random effects analysis for single-measure ICCs was used to determine the applicability of the results to other clinical judges who would be independently rating the measures.

Results

Over 3600 individual measures were anlayzed to determine reliability. Roughly 14% of all measures were not able to be analyzed. Most of these measures were from three patients who were unable to complete the videofluoroscopic protocol because they met "bailout" criteria. Additional measures were unable to be rated because patients refused a particular bolus or because of technical problems, e.g., poor image quality. Intra- and interjudge reliability results for the VFS measures are presented in Tables 2–6. In addition to the missing data points, some correlation values could not be computed, typically because one judge's ratings for the measure were constant. Where a value could not be computed, percent agreement is reported. Kappa values which appear in **bold** type represent reliability that is "moderate" to "almost perfect" in accordance with a scale for kappa values created by Landis and Koch [16]. When values for Kendall's tau or intraclass correlation coefficient (ICC) appear in **bold**, this indicates the measure's reliability was significant at $p < 0.01$.

Penetration–Aspiration

In Table 2, reliability is reported for ratings of penetration–aspiration by two methods: (1) present vs. absent and (2) an 8-point penetration–aspiration scale [13]. Individual and group kappas are provided for intrajudge reliability (column 2 labeled Intra) and interjudge reliability (column 3 labeled Inter) using the present–absent scale. For the 8-point scale, intrajudge reliability (column 4 labeled Intra) is reported using Kendall's tau (τ), and interjudge

Table 2. Reliability for videofluoroscopic measures of penetration–aspiration rated on two scales. Intrajudge reliability results from comparisons of judge 1 ratings made at two different times. Interjudge reliability results from comparisons among all three judges

	Present–absent ratings		8-point scale ratings[b]	
Consistency	Intra[a]	Inter	Intra	Inter
1. Thin liquid (5 cc)	$\boldsymbol{\kappa = 0.843}$	$\kappa = 0.400$	$\tau = 0.467$	ICC = 0.114
2. Thin liquid (5 cc)	$\kappa = 0.757$	$\kappa = 0.274$	$\boldsymbol{\tau = 0.750}$	ICC = 0.380
3. Thin liquid (10 cc)	$\kappa = 0.693$	$\boldsymbol{\kappa = 0.415}$	$\tau = 0.473$	ICC = 0.591
4. Thin liquid (10 cc)	$\kappa = 0.530$	$\kappa = -0.138$	$\tau = 0.355$	ICC = 0.085
5. Thick liquid (5 cc)	$\kappa = 0.755$	$\kappa = 0.067$	$\tau = 0.332$	ICC = 0.647
6. Thick liquid (5 cc)	$\kappa = 0.724$	$\kappa = 0.190$	$\tau = 0.284$	ICC = 0.628
7. Thick liquid (10 cc)	$\kappa = 0.577$	$\kappa = 0.352$	$\tau = 0.172$	ICC = 0.623
8. Thick liquid (10 cc)	$\kappa = 0.886$	$\kappa = -0.200$	100%	ICC = 0.008
9. Puree (5 cc)	$\kappa = 90\%$	$\kappa = 0.194$	83%	ICC = 0.224
10. Puree (5 cc)	$\kappa = 90\%$	$\kappa = 0.194$	83%	ICC = 0.224
11. Solid (1/4 cookie)	$\kappa = 0.893$	$\kappa = -0.081$	100%	ICC = 0.322
12. Solid (1/4 cookie)	$\kappa = 1.000$	$\kappa = -0.269$	100%	ICC = 0.166

[a]Intrajudge reliability for present–absent was analyzed with Cohen's kappa, and interjudge reliability was analyzed with group kappas. Scale for kappa values: below 0.00 = poor agreement; 0.00–0.20 = slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement; 0.81–1.00 = almost perfect agreement. **Bold** type indicates that reliability for the measure was "moderate" to "almost perfect" in agreement or that percent agreement is 90% or better. Percent agreement is reported when kappa could not be computed.

[b]Intrajudge reliability for the 8-point scale was analyzed using Kendall's tau, and interjudge reliability was analyzed using an intraclass correlation coefficient. **Bold** type indicates that the value is significant at $p < 0.01$ or percent agreement is 90% or better. Percent agreement is reported when tau or ICC could not be computed.

Table 3. Reliability for the videofluoroscopic measures of lingual control and oral residue. Intrajudge reliability results from comparisons of judge 1 ratings made at two different times. Interjudge reliability results from comparisons among all three judges

	Lingual control		Oral residue	
Consistency	Intra[a]	Inter[b]	Intra[a]	Inter[b]
1. Thin liquid (5 cc)	$\kappa = 0.308$[c]	$\kappa = 0.030$	$\kappa = 0.217$	$\kappa = 0.125$
2. Thin liquid (10 cc)	$\boldsymbol{\kappa = 0.632}$	$\kappa = 0.364$	$\kappa = 0.759$	$\kappa = 0.253$
3. Thick liquid (5 cc)	$\boldsymbol{\kappa = 1.000}$	$\kappa = 0.077$	$\kappa = 1.000$	$\kappa = 0.478$
4. Thick liquid (10 cc)	$\boldsymbol{\kappa = 1.000}$	$\kappa = 0.022$	$\kappa = 1.000$	$\kappa = -0.071$
5. Puree (5 cc)	$\boldsymbol{\kappa = 0.638}$	$\kappa = 0.009$	$\kappa = 1.000$	$\kappa = 0.333$
6. Solid (1/4 cookie)	82%	27%	**92%**	37%

[a]Intrajudge reliability was analyzed with Cohen's kappa.

[b]Interjudge reliability was analyzed with a group kappa.

[c]Scale for kappa values: below 0.00 = poor agreement; 0.00–0.20 = slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement; 0.81–1.00 = almost perfect agreement. **Bold** type indicates that reliability for the measure was "moderate" to "almost perfect" in agreement or that percent agreement is 90% or better. Percent agreement is reported when kappa could not be computed.

reliability (column 5 labeled Inter) is reported using intraclass correlation coefficients (ICC). Kappas could be computed to determine intrajudge reliability for rating penetration–aspiration as present or absent. Intrajudge reliability on these was "moderate" to "almost perfect." For two puree swallows, kappas could not be computed because of the high number of normal ratings, but percent agreement was 90% for that consistency. "Moderate" interjudge reliability for binary (present–absent) ratings of penetration–aspiration was achieved for only 1 of the 12 consistencies and bolus sizes. When utilizing the 8-point penetration–aspiration scale, 4 of the 12 consistencies and bolus sizes were rated with significant intrajudge reliability, and none were rated with significant ($p < 0.01$) interjudge reliability.

Reliability values for the other VFS measures are reported in Tables 3–5. The first column in each table lists the consistency and bolus size evaluated. Ratings for a particular consistency and bolus size were reported as abnormal if an abnormal rating was given for at least one of the two swallows

Table 4. Reliability for the videofluoroscopic measures of vallecular residue, pyriform residue, and hypopharyngeal residue. Intrajudge reliability results from comparisons of judge 1 ratings made at two different times. Interjudge reliability results from comparisons among all three judges

	Vallecular residue		Pyriform residue		Hypopharyngeal residue	
Consistency	Intra[a]	Inter[b]	Intra[a]	Inter[b]	Intra[a]	Inter[b]
1. Thin liquid (5 cc)	**$\kappa = 0.915$**[c]	$\kappa = 0.355$	**$\kappa = 0.724$**	$\kappa = 0.165$	**$\kappa = 0.898$**	$\kappa = 0.008$
2. Thin liquid (10 cc)	**$\kappa = 0.848$**	$\kappa = 0.309$	**$\kappa = 0.618$**	$\kappa = 0.087$	**$\kappa = 0.766$**	$\kappa = -0.069$
3. Thick liquid (5 cc)	$\kappa = 0.243$	$\kappa = 0.341$	$\kappa = 0.323$	$\kappa = 0.108$	**$\kappa = 0.600$**	$\kappa = 0.154$
4. Thick liquid (10 cc)	**$\kappa = 0.444$**	$\kappa = 0.001$	**$\kappa = 0.444$**	$\kappa = 0.010$	$\kappa = 0.338$	$\kappa = -0.057$
5. Puree (5 cc)	**$\kappa = 0.611$**	$\kappa = 0.325$	$\kappa = 0.276$	$\kappa = -0.161$	**$\kappa = 0.595$**	$\kappa = 0.117$
6. Solid (1/4 cookie)	**$\kappa = 0.553$**	**$\kappa = 0.417$**	**$\kappa = 0.792$**	$\kappa = 0.077$	**$\kappa = 0.440$**	$\kappa = 0.138$

[a]Intrajudge reliability was analyzed with Cohen's kappa.
[b]Interjudge reliability was analyzed with a group kappa.
[c]Scale for kappa values: below 0.00 = poor agreement; 0.00–0.20 = slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement; 0.81–1.00 = almost perfect agreement. **Bold** type indicates that reliability for the measure was "moderate" to "almost perfect" in agreement.

Table 5. Reliability for the videofluoroscopic measures of epiglottic function and hyolaryngeal elevation and percent agreement for cricopharyngeal function. Intrajudge reliability results from comparisons of judge 1 ratings made at two different times. Interjudge reliability results from comparisons of all three judges

	Epiglottic function		Hyolaryngeal elevation		Cricopharyngeal function	
Consistency	Intra[a]	Inter[b]	Intra[a]	Inter[b]	Intra[a]	Inter[b]
1. Thin liquid (5 cc)	**$\kappa = 0.558$**[c]	$\kappa = 0.003$	**$\kappa = 0.573$**	**$\kappa = 0.431$**	**100**	**92**
2. Thin liquid (10 cc)	**$\kappa = 0.592$**	$\kappa = -0.109$	$\kappa = 0.286$	$\kappa = 0.323$	**92**	75
3. Thick liquid (5 cc)	**$\kappa = 0.444$**	80%	$\kappa = 0.242$	$\kappa = 0.286$	**93**	85
4. Thick liquid (10 cc)	**$\kappa = 0.455$**	84%	$\kappa = 0.259$	$\kappa = 0.243$	**93**	75
5. Puree (5 cc)	**$\kappa = 0.600$**	**92%**	**$\kappa = 0.443$**	$\kappa = 0.287$	**94**	80
6. Solid (1/4 cookie)	**$\kappa = 0.435$**	**90%**	**$\kappa = 0.462$**	$\kappa = 0.169$	**100**	**90**

[a]Intrajudge reliability was analyzed with Cohen's kappa.
[b]Interjudge reliability was analyzed with a group kappa.
[c]Scale for kappa values: below 0.00 = poor agreement; 0.00–0.20 = slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement; 0.81–1.00 = almost perfect agreement. **Bold** type indicates that reliability for the measure was "moderate" to "almost perfect" in agreement or that percent agreement is 90% or better. Percent agreement is reported when kappa could not be computed.

administered. Therefore, 6, rather than 12, consistency–bolus size combinations are presented. Columns labeled Intra provide intrajudge kappa values for each consistency and bolus size for the measure being evaluated. Columns labeled Inter provide group kappa ratings.

Lingual Control and Oral Residue

Intra- and interjudge reliability for binary (normal–abnormal) ratings of lingual control and oral residue are provided in Table 3. For lingual control, four of the six kappas for intrajudge reliability were in the range of "moderate" to "almost perfect." For oral residue, four of the six kappas for intrajudge reliability were within the range of "moderate" to "almost perfect," and the ratings for the solid consistency showed 92% agreement. Interjudge reliability was poor for lingual control and oral residue. Only one of the ratings for oral residue (5-cc thick liquid) demonstrated even "moderate" reliability, and none of the ratings for lingual control were rated with sufficient interjudge reliability.

Vallecular, Pyriform, and Hypopharyngeal Residues

Intra- and interjudge reliability for binary (present–absent) ratings of vallecular, pyriform, and

Table 6. Reliability for videofluoroscopic measures of duration. Intrajudge reliability results from comparisons of judge 1 ratings made at two different times. Interjudge reliability results from comparisons among all three judges

Measure and consistency	Intra[a] Range of *r* values[c]	Inter[b] Range of intraclass correlations (ICC)[c]
Oral transit duration		
Thin liquid	0.315–0.765	0.011–0.342
Thick liquid	**0.820–0.934**[d]	−0.080–0.287
Puree	**0.669–0.871**	0.069–0.274
Solid	**0.623–0.774**	0.240–0.450
Pharyngeal transit duration		
Thin liquid	0.019–0.736	−0.045–0.180
Thick liquid	0.343–0.985	−0.205–0.357
Puree	**0.864–0.974**	0.286–0.442
Solid	**0.808–0.996**	0.432–0.462
Total swallow duration		
Thin liquid	0.200–0.997	0.318–0.138
Thick liquid	**0.858–0.987**	0.149–0.391
Puree	**0.935–0.995**	0.268–0.409
Solid	**0.945–0.995**	0.337–0.462
Pharyngeal delay time		
Thin liquid	0.015–0.882	−0.093–0.151
Thick liquid	**0.383–0.998**	−0.231–0.338
Puree	**0.859–0.932**	0.297–0.337
Solid	**0.870–0.996**	0.458–0.459
Duration of UES opening		
Thin liquid	**0.840–0.971**	−0.052–(−)0.014
Thick liquid	0.072–0.676	−0.093–0.197
Puree	−0.180–0.855	0.076–0.098
Solid	−0.360–0.674	−0.090–0.031

[a]Intrajudge reliability was calculated with Pearson's *r*.
[b]Interjudge reliability was calculated with an Intraclass Correlation Coefficient.
[c]Ranges are reported for each consistency because of the high number of ratings per consistency.
[d]**Bold** type indicates that at least 75% of the reliability calculations for that measure and that viscosity were significant at $p < 0.01$.

hypopharyngeal residues are provided in Table 4. For vallecular residue, five of the six kappas for intrajudge reliability were within the range of "moderate" to "almost perfect." Only one consistency, solid, was rated with "moderate" interjudge reliability for vallecular residue. For pyriform residue, four of the six kappas for intrajudge reliability were within the range of "moderate" to "almost perfect." None of the interjudge ratings for pyriform sinus residue were made with even "moderate" agreement. For hypopharyngeal residue, five of the six kappas for intrajudge reliability were within the range of "moderate" to "almost perfect." None of the interjudge ratings for hypopharyngeal residue demonstrated even "moderate" agreement.

Epiglottic Function, Hyolaryngeal Elevation, and Cricopharyngeal Function

Intra- and interjudge reliability for binary (normal–abnormal) ratings of epiglottic function, hyolaryngeal elevation, and cricopharyngeal function are provided in Table 5. For epiglottic function, all six of the kappas for intrajudge reliability indicate "moderate" agreement. Because of the high number of normal ratings, four of the six interjudge kappas could not be computed, but puree and solid were rated with 90% or better agreement. For hyolaryngeal elevation, three of the six kappas for intrajudge reliability were within the range of "moderate" agreement. Only thin liquid (5 cc) indicated at least moderate agreement for interjudge reliability. Because of the high number of normal ratings for cricopharyngeal function, kappas could not be computed for any of the bolus sizes and consistencies used to determine intra- or interjudge reliability. Intrajudge percent agreement was 90% or better for all bolus sizes–consistency combinations. Only two (5-cc thin and solid) were rated with interjudge agreement of 90% or better. Inter- and intrajudge percent agreement values for cricopharyngeal function are probably inflated because of the high number of normal ratings.

Duration Measures

Reliability for duration measures is shown in Table 6. The Intra column provides Pearson's correlations (*r*) for each duration measure and consistency. The Inter column provides intraclass correlation coefficients (ICC) for each duration measure and consistency. Five duration measures were rated with a millisecond timer: (1) oral transit duration (OTD), (2) pharyngeal transit duration (PTD), (3) total swallow duration (TSD), (4) pharyngeal delay time (PDT) [14], and (5) duration of upper esophageal sphincter opening (DUESO). A description of each measure is provided in Appendix A. Reliability judges reviewed each videotape and recorded the times for the beginning and the end of each measure based on the described physiologic markers. The time between markers was calculated. In Table 6, the first column lists the measure and each consistency timed. The second column lists ranges of Pearson's *r* values for each measure and consistency. For intrajudge reliability for four thin liquid boluses and four thick liquid boluses (two 5 cc and two 10 cc of each consistency) were rated at two separate

times by the primary study clinician. The first and second ratings for the four thin liquid swallows and the four thick liquid swallows were correlated for intrajudge reliability. This is the case for all duration measures (OTD, PTD, TSD, PDT, DUESO). Ranges are provided for those correlations to reduce the amount of data presented in the table. Pearson values listed in **bold** type indicate that at least three of the four correlations for thin or thick liquid were significant at $p < 0.01$. For puree and solid boluses, only two swallows each were rated (5-cc puree and 1/4 cookie/solid). Therefore, intrajudge correlations for both swallows rated are shown for puree and solid consistencies.

Intrajudge reliability results are shown in column 3. ICC were calculated to determine reliability among the three judges' timed measures. For thin and thick liquid boluses, four swallows (two 5 cc and two 10 cc of each consistency) were rated by all judges. Correlations were derived for each swallow. Ranges of ICC values are provided to reduce the amount of data presented in the table. ICC values listed in **bold** type indicate that at least three of the four correlations for thin or thick liquid were significant at $p < 0.01$. For puree and solid boluses, only two swallows each were rated (5-cc puree and 1/4 cookie/solid). Therefore, interjudge correlations for both swallows rated are shown for puree and solid consistencies.

Intrajudge reliability for OTD, PTD, TSD, and PDT was significant for most consistencies (thin liquid is the most prominent exception). DUESO was not rated with sufficient intrajudge reliability, as only one consistency (thin liquid) was rated within acceptable limits. Interjudge reliability for duration measures was poor. None of the measures and consistencies were rated within acceptable limits ($p < 0.01$).

Discussion

Intrajudge reliability for VFS measures, to our knowledge, has not been widely reported. Kuhlemeier et al. [3] examined intrajudge reliability for VFS measures and found it to be very similar to interjudge reliability. However, none of the measures were rated using frame-by-frame analysis, and, as the authors indicated, this could reduce intrajudge reliability. They reported that penetration–aspiration, oral residue, and pharyngeal retention were rated reliably. Ratings on other measures that target the specific biomechanical aspects of swallowing—velopharyngeal apposition, laryngeal elevation, epiglottic tilt, and pharyngo-esophageal opening—was less acceptable.

We used frame-by-frame analysis to obtain quantitative data. Nevertheless, our results, using similar measures, are similar to those reported by Kuhlemeier et al. For example, despite the fact that we subdivided pharyngeal retention into three ratings—vallecular residue, pyriform residue, and hypopharyngeal residue—intrajudge reliability was at least "moderate" for most of those ratings. Intrajudge reliability was similarly promising for a judgment of the presence or absence of penetration–aspiration and oral residue. Superficially, our intrajudge reliability results for epiglottic function and cricopharyngeal function were promising; however, high numbers of normal ratings make these results suspect. In addition, Kuhlemeier et al. [3] used nine clinical judges for the analysis of intrajudge reliability. Despite our less substantive sample size (only the ratings of the primary study clinician were used to determine intra rater reliability), our results in general support those of Kuhlemeier et al.

To our knowledge, no previous studies have reported intrajudge reliability for duration measures without pretraining the observer(s) to criterion performance. In this investigation, intrajudge reliability for all duration measures except DUESO approaches acceptability. Intrajudge reliability was high for these measures in previous investigations [11,12], but training to criterion performance was employed. Two points should be kept in mind: First, reliability based on correlational analysis can be fragile because observations can be significantly related but significantly different. Second, our assessment of intrajudge reliability is based on a single judge. Additional studies of intrajudge reliability that use a larger sample of judges are necessary.

Interjudge reliability results in this investigation are largely consistent with interjudge reliability results in other reports that did not pretrain judges to criterion performance. Wilcox et al. [1] found interjudge agreement to be poor, less than 70% for most measures, including the presence or absence of aspiration and oral and pharyngeal residue. Ekberg et al. [2] observed high interjudge reliability only for penetration–aspiration, the presence or absence of Zenker's diverticula, and the presence or absence of pharyngeal constriction. Perlman et al. [5] reported more impressive results, indicating that videofluoroscopic measures were rated for the most part with 70% or higher agreement. However, their judges examined all 330 patients together prior to the reported analysis and reached consensus on every measure for every

swallow. This or similar methods of training to criterion performance may be necessary to achieve acceptable reliability in rating VFS examination measures. This assumption is supported by a recent investigation [4] that studied interjudge reliability under three conditions: (1) judges received no training or conferring, (2) judges discussed VFS studies while rating within a group, and (3) judges rated a final video independently after the discussion which occurred in the second condition. Results indicated that clinicians were essentially unreliable in the "no training or conferring" condition and very reliable when training and conferring occurred. Interjudge reliability appeared to improve in the third condition—independent rating after discussion occurred. Similarly, training to criterion performance appears to enhance interjudge reliability for duration measures. Our judges were not pretrained to criterion performance, and none of the duration measures were rated with acceptable interjudge reliability for the majority of bolus types. In studies that used pretraining to criterion performance for duration measures [11,12], all ratings were made with acceptable interjudge reliability.

Our results suggest two tentative assumptions regarding reliability for VFS examinations. First, intrajudge reliability on measures of penetration–aspiration, lingual function, oral residue, vallecular residue, pyriform sinus residue, and hypopharyngeal residue appears acceptable. Thus, an experienced clinician may employ consistent standards for rating these VFS measures across patients and time. Second, interjudge reliability for most measures, with the exception of a binary rating of aspiration, appears to vary among clinicians and is unacceptable. However, as demonstrated in investigations that provided pretraining to criterion performance [4–6,11–13], clinicians can be trained to achieve acceptable interjudge reliability.

It appears that additional work is necessary to establish both intra- and interjudge reliability for commonly used VFS measures. While our intrajudge reliability results appear promising, they are based on a single judge. Moreover, we did not examine the validity of that judge's observations. Achieving acceptable interjudge reliability probably will require training observers to a criterion performance. The nature and extent of that training needs to be established. If the VFS examination is used to make a diagnosis, direct management, and evaluate outcome, then obviously clinician observations on the measures used in the examination must be reliable.

Appendix A: Videofluoroscopic Measures Evaluated for Reliability

Measure	Description
Measures of Oropharyngeal Function	
Penetration–Aspiration[a]	Present if material enters into the laryngeal vestibule; absent if no material enters the laryngeal vestibule[b]
Reduced lingual control	Present if evidence of reduced lingual propulsion of bolus
Oral residue	Present if more than trace coating; absent if trace or no material present
Vallecular residue	Present if more than trace coating; absent if trace or no material present
Pyriform residue	Present if more than trace coating; absent if trace or no material present
Hypopharyngeal residue	Present if more than trace coating; absent if trace or no material present
Epiglottis dysfunction	Epiglottis does not invert completely
Reduced hyolaryngeal excursion	Excursion is reduced or absent
Cricopharyngeal prominence	Evidence of prominence as bolus passes through UES
Duration Measures[c]	
Oral transit duration	Beginning of posterior movement of the bolus to the arrival of the bolus head at ramus of mandible
Pharyngeal transit duration	From bolus head at ramus of mandible to bolus head entering cricopharyngeus
Total swallow duration	Oral transit duration plus pharyngeal transit duration
Pharyngeal delay time	From the bolus head arrival at the point where the lower rim of the mandible crosses the tongue base until the first laryngeal elevation
Duration of UES[d] opening	From the time the UES opens to the time UES closes

[a]All measures other than durations were rated in a binary manner: normal–abnormal or present–absent.
[b]Penetration–aspiration was also rated with an 8-point penetration–aspiration scale [13].
[c]All duration measures were analyzed in milliseconds.
[d]UES = upper esophageal sphincter.

References

1. Wilcox F, Liss JM, Siegel GM: Interjudge agreement in videofluoroscopic studies of swallowing. *J Speech Hear Res 39:*144–152, 1996
2. Ekberg O, Nylander G, Fork FT, Sjoberg S, Birch-Iensen M, Hillarp B: Interobserver variability in cineradiographic assessment of pharyngeal function during swallow. *Dysphagia 3:*46–48, 1988
3. Kuhlemeier KV, Yates P, Palmer JB: Intra- and interrater variation in the evaluation of videofluorographic swallowing studies. *Dysphagia 13:*142–147, 1998

4. Scott A, Perry A, Bench J: A study of interrater reliability when using videofluoroscopy as an assessment of swallowing. *Dysphagia 13:*223–227, 1998
5. Perlman AL, Booth BM, Grayhack JP: Videofluoroscopic predictors of aspiration in patients with oropharyngeal dysphagia. *Dysphagia 9:*90–95, 1994
6. Smith CH, Logemann JA, Colangelo LA, Rademaker AW, Pauloski BR: Incidence and patient characteristics associated with silent aspiration in the acute care setting. *Dysphagia 14:*1–7, 1999
7. Tracy JF, Logemann JA, Kahrilas PJ, Jacob P, Kobara M, Krugler C: Preliminary observations on the effects of age on oropharyngeal deglutition. *Dysphagia 4:*90–94, 1989
8. Logemann JA, Pauloski BR, Colangelo L, Fujiu M, Kahrilas PJ: Effects of sour bolus on oropharyngeal swallowing measures in patients with neurogenic dysphagia. *J Speech Hear Res 38:*556–563, 1995
9. Daniels SK, McAdam CP, Brailey K, Foundas AL: Clinical assessment of swallowing and prediction of dysphagia severity. *Am J Speech Lang Pathol 6(4):*17–23, 1997
10. McCullough GH, Wertz RT, Rosenbek JC, Dinneen C: Clinicians' preferences and practices for clinical/bedside and videofluoroscopic examinations of swallowing in adults. *Am J Speech Lang Pathol 8:*149–163, 1999
11. Lof G, Robbins J: Test–retest variability in normal swallowing. *Dysphagia 4:*236–242, 1990
12. Robbins JA, Hamilton J, Lof GL, Kempster GB: Oropharyngeal swallowing in normal adults of different ages. *Gastroenterology 103:*823–829, 1992
13. Rosenbek JC, Robbins J, Roecker EB, Coyle JL, Wood JL: A penetration–aspiration scale. *Dysphagia 11:*93–98, 1996
14. Logemann JA, Shanahan T, Rademaker AW, Kahrilas PJ, Lazar R, Halper A: Oropharyngeal swallowing after stroke in the left basal ganglion/internal capsule. *Dysphagia 8:*230–234, 1993
15. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. *Psychol Bull 86(2):*420–428, 1979