Inter-and Intrajudge Reliability of a Clinical Examination of Swallowing in Adults

Authors
Gary H. McCullough, Robert T. Wertz, John C. Rosenbek, Russell H. Mills,
Katherine B. Ross and John R. Ashford

Abstract

This study investigates inter-and intrajudge reliability of a clinical examination of swallowing in adults. Several investigations have sought correlations between clinical indicators of dysphagia and the actual presence of dysphagia as determined by videofluoros- copy. Whereas some investigations have reported inter- judge reliability for the videofluoroscopic measures employed, none have reported reliability for clinical measures. Without established reliability for rating clinical measures, conclusions drawn regarding the utility of a measure for detecting aspiration can be called into question. Results of the present study indicate that fewer than 50% of the measures clinicians typically employ are rated with sufficient inter- and intrajudge reliability. Measures of vocal quality and oral motor function were rated more reliably than were history measures or measures taken during trial swallows. There is a need to define more clearly the measures employed in clinical examinations and to be consistent in reporting reliability for clinical measures of swallowing function in future research

Gary H. McCullough, Robert T. Wertz, John C. Rosenbek, Russell H. Mills, Katherine B. Ross and John R. Ashford (2000) "Inter-And Intrajudge Reliability Of A Clinical Examination Of Swallowing In Adults" *Dysphagia* #15: pp.58–67 Version of Record available @At www.springer.com [DOI: 10.1007/s004550010002]

Inter-And Intrajudge Reliability Of A Clinical Examination Of Swallowing In Adults

A recent survey [1] of clinicians working with dysphagic patients investigated their preferences and practices for conducting clinical/bedside and videofluoroscopic (VFS) evaluations of swallowing. Results indicated that clinicians differ with regard to which clinical/bedside methods and measures they believe should be employed. Comparisons of methods and measures clinicians prefer and use with methods and measures that have research support indicated that research support is lacking for many of the measures clinicians employ. Furthermore, descriptions of how to elicit and rate measures are few and far between. For example, although several investigations have linked the presence of an "abnormal volitional cough" to the presence of aspiration in stroke patients [2–5], no clear descriptions exist for how to rate an "abnormal volitional cough." Moreover, none of those investigations have reported inter- or intrajudge reliability for rating clinical/bedside measures. It cannot be assumed that different speech-language pathologists are reliable in obtaining history information from charts or in evaluating oral motor, voice, or trial swallow components of an examination. Likewise, it cannot be assumed that one clinician would rate the same patient similarly in a subsequent evaluation. Correlations between clinical signs and VFS signs of aspiration may be spurious unless clear definitions for clinical measures exist and reliability for rating those measures has been demonstrated. Although a number of studies have attempted to explore the relation between specific clinical/bedside measures and actual swallowing function, as judged by VFS examination [2–5], there are no investigations, to our knowledge, that have addressed the reliability of administering and analyzing a clinical/bedside examination of swallowing.

Reliability for rating some of the measures employed in a clinical/bedside examination have been re- ported outside the context of swallowing evaluations. Some of these include the presence of dysarthria [6–8], perception of intelligibility [9–12], and various aspects of vocal quality [13–17]. However, none of these reliability data were gathered within the context of a clinical/bedside swallowing examination; and, to our knowledge, no other reliability data for clinical/bedside measures exist. The purpose of the present study was to examine the inter- and intrajudge reliability of clinical/ bedside examination measures commonly used to assess swallowing function.

Methods

Subjects

Twenty subjects from the Veteran's Administration Medical Center (VAMC) in Nashville, Tennessee, Vanderbilt University Medical Center, and the VAMC in Murfreesboro, Tennessee participated in this study. The inclusion criterion was that the subject suffered a stroke within 6 weeks of the time of examination (almost all were within 2 weeks postonset). Patients with previous strokes were included as long as no swallowing problems were reported to exist from the prior stroke. Exclusion criteria were an anatomical/structural deviation that would affect swallowing and a tracheostomy.

Descriptive data on subjects are located in Table 1. The mean age for subjects was 67.8, and the mean number of days postonset was 7. Locations of the subjects' lesions varied throughout cortical and subcortical areas but were predominantly unilateral. No brainsteam lesions occurred in this sample. VFS examination of each subject showed that 14 had some penetration or aspiration on at least one swallow and that 11 had penetration or aspiration on more than one swallow. Nine swallows were elicited from each subject.

Table 1. Descriptive data of subjects

Subject	Age/sex	DPO[a]	Stroke localization
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 (unspecified)
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

DPO, days postonset; L, left; R, right; MCA, middle cerebral artery; Hx, history.

Design

Three clinicians, all certified speech pathologists with at least 200 hr of experience in dysphagia evaluation and management performed a clinical/bedside examination on each subject. On the first day, the principal investigator (judge 1) administered a clinical/bedside examination with one of the two other clinicians (judge 2 or judge 3). On the following day, judge 1 administered a clinical/bedside examination to the same subject with the other clinician (judge 2 or judge 3). This method of administration was chosen for several reasons. First, intrajudge reliability was considered to be less biased if testing occurred on consecutive days rather than on the same day. Second, this methodology was designed to determine whether a minimal increase (1 day) in postonset time affects intrajudge and interjudge reliability. If an increase in time has an effect on subject performance, interjudge reliability between the judges evaluating the subject on the same day would be better than intrajudge reliability as measured on 2 different days. This is poten- tially confounded by different judges on different days—and, perhaps, poor intrajudge reliability. However, this was deemed the most effective method of obtaining both inter- and intrajudge reliability while examining the effects of time. Judge 2 and judge 3 were randomly assigned to day 1 or day 2 to ensure that reliability was not dependent on the day of the examination.

The clinical/bedside measures employed were those that clinicians believe are important and use in their practice, as indicated by a previous survey [1]. There were four parts to the clinical/bedside ex- amination: history, oral motor, voice, and trial swallows. Measures rated for reliability are listed in Table 2. All ratings were reported in a binary manner (+/− or normal/abnormal). If the patient could not be assessed on any task, the clinician circled *CNA* (cannot assess) on the response form. For the history portion of the examination, each clinician obtained the information separately from the medical chart, patient, family, physician, or nurse, depending on the question. For the oral motor and voice portions of the clinical/bedside examination, judge 1 elicited all responses from the patient to reduce examiner variability, except for the items *tongue strength* and *jaw strength*, be- cause those measures required each clinician to examine the subject physically. Both clinicians, on each day, recorded their observations independently without discussion. Voice measurements were taken by two methods: speech sample and sustained phonation. For the trial swallows portion of the clinical/bedside examination, two swallows of each consistency—thin liquid, thick liquid, puree, and solid—were administered in 5-cc boluses. Ratings of measures involving solid consistencies, however, are not reported because of the high number of normal ratings with that consistency. Statistical analyses could not be performed and percentage of agreement was misleadingly high. Thin and thick liquids were administered from a cup; puree and solids were administered from a spoon. Each clinician administered one of the swallows for each consistency so that he/she could use the four-finger method [18] to judge laryngeal elevation, timing of the initiation of the swallow, the number of swallows per bolus, and the total swallow duration. Thus, each clinician made judgments on those four measures once for each consistency. All other measures were judged with two swallows of each consistency because the measurements could be made by observing the patient's response on the task elicited by either clinician.

No training to criterion occurred before this investigation. Clinicians were provided only with a list of anchors on which to base their normal or abnormal judgments. Training to criterion was avoided be-

Table 2. Clinical/bedside measures evaluated for reliability

History[a]	
Patient reports problem	Presence of feeding tube
Family reports problem	Requires suctioning
Nurse reports problem	Poor oral hygiene
History or current pneumonia	Decreased mental status/dementia
Gastrointestinal disorder	Decreased level of consciousness
Previous head/neck/heart/gastrointestinal surgery	COPD/decreased pulmonary clearance
Other related disease	
Medications	
Oral motor	
Tongue	Palatal gag reflex
Strength—protrusion against resistance	Strength—response to tongue-depressor contact, left and right
Range of motion—side to side movement	
Lips	Pharyngeal gag reflex
Strength—maintain seal against resistance	Strength—response to cotton-tip applicator contact, left and right
Range of motion—pucker and retract	Oral apraxia—pretend to blow out a match, cluck tongue, whistle
Jaw	Dysarthria (from speech sample)
Strength—open and close against resistance	Intelligibility (from speech sample)
Range of motion—side to side movement	Secretion management—appearance of drooling or continual coughing and wet voice quality
Soft palate	
Strength—movement in repeated “Ahs”	
Range of motion—symmetry in same task	
Volitional cough Strength—intensity of cough Quality—wet or dry sound	
Ability to attend to all tasks	
Voice	
From speech sample	From sustained phonation
Dysphonia	Dysphonia
Breathy	Breathy
Harsh	Harsh
Strained/strangled	Strained/strangled
Wet/gurgly	Wet/gurgly
Trial swallows[c]	
Using the four-finger method,[b] is there	Estimate of penetration/aspiration
Delayed swallow (>1 sec)	“Do you believe the person had laryngeal penetration/aspiration?”
Prolonged total swallow duration (>2 sec)	
Decreased laryngeal elevation	Presence of oral stasis after the swallow
Spontaneous cough during or after swallow	3-oz. swallow test—wet voice or coughing up to 1 min
Number of swallows per bolus	Overall swallowing function
Wet voice after the swallow (when saying *ah*)	

[a] All measures rated as present/absent or normal/abnormal.
[b] From Logemann [18].
[c] For thin/thick liquid/pureed consistencies.

cause practicing clinicians, unlike researchers, typically have not received this training. This investigation should, therefore, provide clinically applicable reliability data.

Analysis of Data

For all ratings, Cohen’s kappa coefficients were used for analysis. This statistic was chosen because ratings were made in a binary manner, and the statistic corrects for chance occurrence of significance. However, because of the large number of kappa values employed, chance occurrence of significance was considered possible. To account for this, one could either employ a Bonferroni adjustment or adjust the level of significance to make the criterion for significance more rigid. A Bonferroni adjustment was deemed too severe because kappa values do correct, to some extent, for chance. Therefore, the level of significance was adjusted from 0.05 to 0.01 to ensure that clinical/bedside measures with significant kappa values were reliable. All kappa values are listed in the table unless they could not be computed. Kappa values could not be computed when at least one judge rated all subjects as “normal” for that measure.

Results

History Measures

Table 3 provides reliability data for history measures. Each measure investigated is listed in column 1. Intrajudge reliability (column 2) was not obtained for history measures because most of the measures were obtained from medical records. The clinician responsible for intrajudge ratings was able to refer to the same pages of the medical record on two separate occasions, providing a misleading significance for reliability data. Therefore, interjudge reliability was believed to be of more value for history measures. Columns 3 and 4 provide inter- judge reliability data. Column 3 provides kappa values for paired comparisons between two judges rating the measures on day 1. Column 4 provides kappa values for paired comparisons between two judges rating the measures on day 2. Kappa values appearing in boldface type were statistically significant at the 0.01 level, indicating good reliability. In addition to the measures listed in Table 3, the following history information was obtained from each patient: presence of neurologic insult, presence of a tracheostomy tube, and presence of structural deficit. These measures were not analyzed for reliability because the first was an inclusion criterion for the study and the latter two were exclusion criteria. The reliability for these measures has not been established.

Only five of the 14 measures analyzed for reliability were rated with significant ($p < 0.01$) interjudge reliability on both days: (a) history of pneumonia, (b) gastrointestinal disorder, (c) medications, (d) presence of a feeding tube, and (e) patient requires suctioning. Medications were rated as present or absent based on information in the patient's chart. Reports of individual medications, which change on a regular basis for many patients, were not analyzed for reliability. Most hospitalized patients are on some types of medications. Thus, the reliability of *medications* is of little functional utility.

Table 3. Reliability for history measures; Cohen's kappa were used for each paired comparison

		Intrajudge[b]	
Measure	Intrajudge[a]	Day 1	Day 2
Patient reports problem	N/A	**0.435**	**0.294**[c]
Family reports problem		**0.554**	−0.193
Nurse reports problem		0.027	0.155
History of pneumonia		**0.875**	**0.875**
Gastrointestinal disorder		**0.565**	**0.737**
Previous surgery		0.394	0.271
Related disease		0.231	0.444
Medications		**1.000**	—[d]
Feeding tube		**1.000**	**1.000**
Requires suction		**1.000**	**1.000**
Oral hygiene		0.158	0.200
Poor mental status		0.306	0.281
Decreased consciousness		**0.643**	0.333
COPD[e]		**0.765**	0.077

[a] Intrajudge reliability was not assessed for history measures.
[b] Results are from a comparison of ratings made by judge 1 and ratings made by judge 2 on half the patients and by judge 3 on half the patients.
[c] Boldface type indicates that the reliability for the measure was significant at the 0.01 level.
[d] Measure was not calculable with the kappa statistic.
[e] Chronic obstructive pulmonary disease.

Oral Motor Measures

Table 4 provides reliability data for oral motor measures. Each measure investigated is listed in column 1. Intrajudge reliability kappas are provided in column 2. Column 3 provides kappa values for paired comparisons between two judges rating the measures on day 1. Column 4 provides kappa values for paired comparisons between two judges rating the measures on day 2. Kappa values appearing in boldface type were statistically significant at the 0.01 level, indicating good reliability.

Sixteen of the 19 oral motor measures were rated with significant ($p < 0.01$) intrajudge reliability. Only palatal gag strength (left and right) and pharyngeal gag strength (right) were not rated with sufficient intrajudge reliability. Only 11 of the 19 measures, however, were rated with significant interjudge reliability.

When looking at all ratings on both days, 11 of the 19 oral motor measures were rated with significant ($p < 0.01$) inter- and intrajudge reliability: (a) tongue strength, (b) tongue range of motion, (c) lip strength, (d) lip range of motion, (e) jaw range of motion, (f) volitional cough strength, (g) volitional cough quality, (h) left pharyngeal gag, (i) dysarthria, (j) intelligibility, and (k) management of secretions.

Voice Measures

Table 5 provides reliability data for voice measures. Each measure investigated is listed in column 1. Voice ratings were elicited by two methods: speech sample and sustained phonation. Ratings were made separately for each method. Column 2 provides kappa values for intrajudge reliability ratings. Column 3 provides kappa values for paired comparisons between two judges rating the measures on day 1. Column 4 provides kappa values for paired comparisons between two judges rating the measures on day 2. Kappa values appearing in boldface type were statistically significant at the 0.01 level, indicating good reliability.

All five of the voice measures rated from a speech sample were rated with significant intrajudge reliability. When listening to a sustained *ah*, however, only three of the five measures were rated reliably: dysphonia, breathy, and wet/gurgly.

Four of the five voice measures rated from a speech sample were rated with significant interjudge reliability. Only a measure of strained/strangled quality

Table 4. Reliability for oral motor measures; Cohen's kappa was used for each paired comparison

Measure	Intrajudge[a]	Interjudge[b]	
		Day 1	Day 2
Tongue			
Strength	**0.913**	**0.627**	**0.554**
Range of motion	**1.000**	**0.806**	**0.907**
Lips			
Strength	**0.810**	**0.583**	**0.444**
Range of motion	**0.717**	**0.517**	**0.631**
Jaw			
Strength	**0.459**	**1.000**	–0.053[c]
Range of motion	**0.712**	**0.902**	**0.444**
Soft palate			
Strength	**0.512**	0.234	0.286
Range of motion	**0.602**	0.380	0.389
Volitional cough			
Strength	**0.918**	**0.677**	**0.839**
Quality (wet/dry)	**0.654**	**0.602**	**0.404**
Palatal gag strength			
Left	0.298	**0.633**	0.397
Right	0.113	0.391	0.113
Pharyngeal gag strength			
Left	**0.555**	**0.492**	**0.658**
Right	0.382	**0.506**	0.214
Oral apraxia	**0.618**	**0.727**	0.300
Dysarthria	**0.922**	**0.767**	**0.611**
Intelligibility	**0.757**	**0.770**	**0.404**
Attends to tasks	**0.497**	**0.714**	0.190
Manages secretions	**0.895**	**0.567**	**0.779**

[a]Results are from a comparison of ratings made by judge 1 on 2 consecutive days.
[b]Results are from a comparison of ratings made by judge 1 and ratings made by judge 2 on half the patients and by judge 3 on half the patients.
[c]Boldface type indicates that the reliability for the measure was significant at the 0.01 level.

Table 5. Reliability for voice measures; Cohen's kappa was used for each paired comparison

Measure	Intrajudge[a]	Interjudge[b]	
		Day 1	Day 2
From speech			
Dysphonia	**0.695**	**0.742**	**0.671**
Breathy	**0.829**	**0.685**	**0.362**
Harsh	**0.808**	**0.680**	**0.611**
Strained/strangled	**0.673**	0.338[c]	—[d]
Wet/gurgly	**0.759**	**0.602**	**0.686**
From sustained *ah*			
Dysphonia	**0.886**	**0.786**	**0.550**
Breathy	**0.612**	0.550	0.556
Harsh	0.089	0.491	0.135
Strained/strangled	0.485	0.260	0.131
Wet/gurgly	**0.462**	**0.619**	**0.641**

[a]Results are from a comparison of ratings made by judge 1 on 2 consecutive days.
[b]Results are from a comparison of ratings made by judge 1 and ratings made by judge 2 on half the patients and by judge 3 on half the patients.
[c]Boldface type indicates that the reliability for the measure was significant at the 0.01 level.
[d]Measure was not calculable with the kappa statistic.

was not rated with sufficient interjudge reliability. When rating from a sustained *ah*, only two of the measures (an overall rating of dysphonia and wet/gurgly quality) were rated with significant interjudge reliability.

When looking at all ratings on both days, four of the five voice measures rated from a speech sample were rated with significant intrajudge and interjudge reliability: dysphonia (overall judgment), breathy, harsh, and wet/gurgly. Only a judgment of strained/strangled quality was not rated reliably. When using sustained phonation, only two of the five measures were rated with significant inter- and intrajudge reliability: dysphonia (over- all judgment) and wet/gurgly quality.

Trial Swallow Measures

Table 6 provides reliability data for trial swallow measures. Each measure investigated is listed in column 1. Column 2 provides kappa values for intrajudge reliability ratings made by judge 1 on 2 separate days. Column 3 provides kappa values for paired comparisons between two different judges rating the measures on day 1. Column 4 provides kappa values for paired comparisons between two different judges rating the measures on day 2. Kappa values appearing in boldface type were statistically significant at the 0.01 level, indicating good reliability.

Ten measures were rated. The first eight of those measures were made for three different consistencies: thin liquid, thick liquid, and puree. Consequently, there were 26 measures with intra- and interjudge kappa values. Only seven of those measures were rated with sufficient intrajudge reliability: delayed swallow for thin liquid, total swallow duration for thin liquid, total swallow duration for puree, laryngeal elevation for thin liquid, an estimate of oral stasis, the 3-oz. swallow test, and an overall rating of dysphagia.

Five measures were rated with sufficient interjudge reliability: delayed swallow on thick liquid, total swallow duration on thin liquid, spontaneous cough on thick liquid, the 3-oz. swallow, and an overall rating of dysphagia.

When looking at all ratings on both days, only four of those 26 measures were rated with sufficient intra- and interjudge reliability: an estimate of total swallow duration for thin liquid, an estimate of oral stasis, the 3-oz. swallow test, and an overall rating of dysphagia as normal or abnormal. Reliability for estimating oral stasis should be considered with special caution because the

Table 6. Reliability for trial swallows; Cohen's kappa was used for each paired comparison

Measure	Intrajudge[a]	Interjudge[b]	
		Day 1	Day 2
Delayed swallow			
Thin liquid	**0.658**	0.111	0.174
Thick liquid	0.346	**0.609**	**0.765**
Puree	–0.098	0.446	–0.207
Total swallow duration			
Thin liquid	**0.609**	**0.609**	**0.612**
Thick liquid	**0.452**	**0.852**	0.557
Puree	**0.634**	0.557	0.417
Laryngeal elevation			
Thin liquid	**0.640**	0.200	**0.771**
Thick liquid	–0.085	–0.091	**1.000**
Puree	—[d]	0.067	—[d]
Spontaneous cough			
Thin liquid	0.360	**0.898**	0.360
Thick liquid	**0.452**	**0.771**	**0.638**
Puree	—[d]	**0.767**	—[d]
Swallows/bolus			
Thin liquid	—[d]	—[d]	—[d]
Thick liquid	0.429	0.433	0.429
Puree	—[d]	—[d]	—[d]
Wet voice after swallow			
Thin liquid	0.429[c]	**0.880**	0.467
Thick liquid	–0.105	0.038	–0.190
Puree	—[d]	**0.815**	**0.629**
Penetration/aspiration			
Thin liquid	0.374	**0.604**	0.469
Thick liquid	0.038	0.362	0.595
Puree	–0.207	0.598	—[d]
Oral stasis[e]			
3-oz. swallow	**0.436**	**0.858**	**0.438**
Dysphagia (overall rating)	**0.596**	**0.728**	**0.685**

[a] Results are from a comparison of ratings made by judge 1 on 2 consecutive days.

[b] Results are from a comparison of ratings made by judge 1 and judge 2 on half the patients and by judge 3 on half the patients.

[c] Boldface type indicates that the reliability for the measure was significant at the 0.01 level.

[d] Measure was not calculable with the kappa statistic.

[e] Because of the high numbers of normal ratings, kappas were not calculable for oral stasis, but agreement was 100% or within 1 of 100% for all consistencies.

Table 7. Clinical/bedside measures with significant inter- and intrajudge reliability

- History[a]
  - History or current pneumonia
  - Gastrointestinal disorder
  - Medications
  - Presence of feeding tube
  - Requires suctioning
- Oral motor
  - Tongue
    - Strength—protrusion against resistance
    - Range of motion—side to side movement
  - Lips
    - Strength—maintain seal against resistance
    - Range of motion—pucker and retract
  - Volitional cough Strength—
    - intensity of cough Quality—
    - wet or dry sound
  - Dysarthria (from speech sample)
  - Intelligibility (from speech sample)
  - Secretion management—appearance of drooling or continual coughing and wet voice quality
- Voice
  - From speech sample
    - Dysphonia
    - Breathy
    - Harsh
    - Wet/gurgly
  - From sustained phonation
    - Dysphonia
    - Wet/gurgly
- Trial swallows[b]
  - Prolonged total swallow duration (four-finger method)
  - Presence of oral stasis after the swallow
  - 3-oz. swallow test—wet voice or coughing up to 1 min after the swallow
  - Overall swallowing function

[a] Measures rated as present/absent or normal/abnormal.

[b] For thin and thick liquid and pureed consistencies.

high number of normal ratings made statistical analysis of reliability impossible.

Reliable Clinical/Bedside Measures

Table 7 provides a list of measures that were rated with significant inter- and intrajudge reliability in this investigation. Measures must have significant intrajudge reliability to demonstrate that one's own ratings have an internal standard for consistency. Measures also must be reliable between judges to demonstrate that one's own internal standards are similar to the standards of other clinicians. Without consistency between clinicians, no comparisons can be made between patients seen by different clinicians, and data for samples of different patient populations would be meaningless.

Overall, 24 of the 54 clinical/bedside measures (44%) were rated with sufficient inter- and intrajudge reliability. Five of 18 history measures (28%) were obtained reliably. Eleven of 19 oral motor measures (58%) were obtained reliably. Six of 10 voice measures (60%) were obtained reliably. And, when considering each consistency separately for each measure, only four of the 26 (15%) were rated reliably. Thus, oral motor and voice measures were the most reliably obtained, as shown in Figure 1.

Discussion

Results of this study indicate areas of relative strength and weakness in speech-language pathologists' abilities

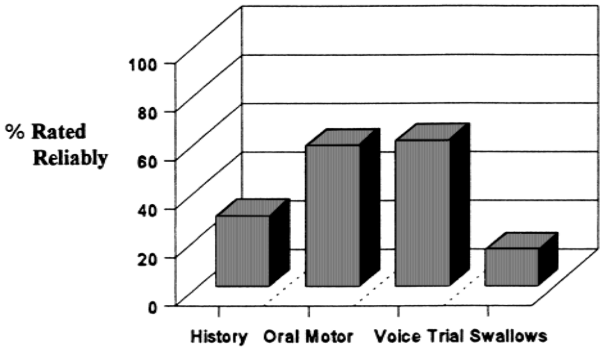

Fig. 1. The percentage of measures with significant inter- and intrajudge reliability within each major section of the clinical/bedside examination of swallowing.

to make reliable judgments on clinical/bedside measures that may relate to swallowing function.

History

The most *un*reliable history components were those requiring clinicians to obtain verbal information from someone—patient, family, or nurse (Table 3). Different nurses are available at different times; family members are present at one time and not another; and the patient, who may have language or cognitive deficits, is not al- ways a reliable historian. A patient's report of swallowing capability did not correlate well with actual swallowing function in a study that examined that measure [19]. The most reliable judgments were visual judgments: the presence of a tracheostomy tube (which was an exclusion criterion); the presence of a nasogastric tube, jejunostomy tube, or percutaneous endoscopic gastrostomy tube; and the presence of used suctioning equipment in the room ("patient requires suctioning"). These measures have received research support for their inclusion in swallowing evaluation. The presence of mechanical devices has been observed to correlate highly with aspiration [20] or other, more long-term complications in several studies [21–29]. The presence of used suctioning equipment relates to management of secretions, which has been demonstrated to have a relation to aspiration pneumonia [30–31]. Reliability of other information obtained from charts such as gastrointestinal history, chronic obstructive pulmonary disease (COPD), and medications varied. This could depend on how easy the information was to locate in the chart. Reasons for low reliability on some of those measures, however, is unknown.

One should also consider reliability of measures in conjunction with their assumed importance for detecting a swallowing problem. The majority of clinicians surveyed in a recent investigation [1] reported that all of the history measures investigated for reliability in this study were either "important" or "essential" to obtain. Other investigations have offered support for the inclusion of many of these measures. Some items that have research support for their relation to either aspiration or aspiration pneumonia, however, were not obtained reliably in this study. These were decreased mental status [31], decreased level of consciousness [32–33], and the presence of COPD/decreased pulmonary clearance [30,34]. Low reliability questions the benefit of making these measures in an attempt to establish a risk protocol for dysphagia. Better anchors, or perhaps more well- defined and available medical diagnoses, for these items may be needed. The presence or history of aspiration pneumonia has also received research support [24]. Reliability for obtaining that information in the present study was high.

Oral Motor

Similar to history measures, visual judgments resulted in the best reliability for oral motor ratings (Table 4). The more easily viewed muscles, structures, and functions were the ones for reliably rated (i.e., lips, tongue, and jaw). The only exception to this was jaw strength, whose interjudge reliability was high for day 1 paired comparisons but low for day 2 paired comparisons.

The only tasks with low *intra*judge reliability were judgments of palatal and pharyngeal gag. The sensitivity of these measures for detecting dysphagia have been reported with mixed results in recent literature [29,35–37]. For unknown reasons, judging a left pharyngeal gag had lower reliability than judging a right pharyngeal gag. When combining the judgments of a *left* pharyngeal gag and a *right* pharyngeal gag into one judgment of pharyngeal gag, reliability is high. The utility of combining left and right pharyngeal gag information versus separating them is unknown, as is the utility in using the measure at all. We also cannot be certain why better reliability was achieved with a pharyngeal gag than with a palatal gag. One could speculate that the pharyngeal gag provokes a more obvious response than a palatal gag, but this cannot be clearly discerned from this study. In addition, of all the measures listed in Table 4, only palatal gag and pharyngeal gag were not believed to be important or essential by the majority of clinicians surveyed [1]. Lack of confidence in using gag reflexes may stem from mixed results in research. With questionable reliability, future investigations of palatal and pharyngeal gags would benefit from reporting reliability and from specific instructions on elicitation and rating of responses.

Measurements of dysarthria and speech intelligibility achieved high inter- and intrajudge reliability. Dysarthria has received support in data-based research as a clinical indicator for aspiration [4,38–39]. Likewise, dysarthria is believed by surveyed clinicians to be an important part of a clinical/bedside assessment of swallow- ing [1]. However, it must be recognized that aspiration may exist unaccompanied by dysarthria in adults.

Voice

Four of the five voice ratings were made with high inter- and intrajudge reliability when using a speech task (Table 5). Using a sustained *ah* task, however, reduced the reliability on several judgments and did not add valuable information to the overall assessment, indicating that employing this task in a clinical/bedside screen may be more time consuming than useful. One might speculate that sustaining *ah* is unnatural and provokes strain in patients who have, typically, normal voicing. Having the patient generate a short speech sample—a description of "the cookie thief" picture in this study—appears to be the most reliable method of rating vocal quality. For patients with visual deficits or other deficits that prevent them from providing a description of a picture, spontaneous speech may be elicited in conversation, such as a discussion of one's present or former occupation.

The high reliability obtained in judging vocal quality may be important. Previous studies have demonstrated significant correlations between the presence of dysphonia and the presence of aspiration or aspiration pneumonia [3,4,30,37]. No reliability data were reported in these studies, but the potential for dysphonia to be a reliable, sensitive, and specific sign for detecting aspiration exists, especially in light of the strong reliability observed without training to criterion in the present investigation.

Trial Swallows

Providing the patient with different food and liquid consistencies and rating them on different swallowing durations, laryngeal elevation, wet voice after the swallow, spontaneous cough, and oral stasis is a common practice in the clinical/bedside swallowing evaluation [1]. With the $p < 0.01$ significance level, very few trial swallow measures were rated with acceptable inter- and intra- judge reliability. Those measures meeting this signifi- cance criterion were total swallow duration for thin liq- uid, the presence of oral stasis, coughing or wet voice on the 3-oz. swallow test, and an overall rating of dysphagia (Table 6). As noted in the Results section, data for oral stasis are questionable.

Although low reliability for trial swallow measurements is alarming, certain factors need to be considered. Intrajudge reliability for these measures could be low because of variability in swallowing function. Trial swallows may change from one day to the next as the patient's health status improves or deteriorates. Fluctuation in swallowing is apt to differ widely, not only from day to day but also from swallow to swallow. Therefore, measures such as coughing, wet voice, and penetration/aspiration could easily produce low intrajudge reliability despite consistency in rating procedures. Therefore, low intrajudge reliability when coupled with high interjudge reliability could indicate that the variability exists in the patient and not in the clinical judge.

Low interjudge reliability also could occur because of variability in the patient from swallow to swallow or from variability in the examiners with the examination method employed. For half of the trial swallow measurements—delayed swallow, total swallow duration, laryngeal elevation, and number of swallows per bolus—judgments were made with the four-finger method [18]. In this method, four fingers are placed on the throat: under the chin, on the hyoid, and on the top and bottom of the thyroid cartilage. Only one clinician at a time can make these judgments. Therefore, judgments of these measures were made on two separate, consecutive swallows. Lof and Robbins [40] demonstrated that marked variability exists within subjects from swallow to swallow. Thus, patient behavior could have changed or clinicians could have been responsible for the variability when attempting to use the four-finger method; the answer remains unknown. What this investigation does demonstrate is that intra- and interjudge reliability for the four-finger method is sporadic and requires further investigation.

Use of the 3-oz. swallow has been demonstrated, in at least one study, to be a reliable detector of penetration/aspiration [41]. Intra- and interjudge reliabilities for this measure were high in the present investigation. There is no clear reason why reliability for spontaneous cough and wet voice after the swallow were low when reliability for a measure that uses both of those judgments together (3-oz. swallow test) was high. Our best explanation is that the 3-oz. swallow provides an either/ or judgment; thus, only one negative response has to be detected to produce a negative result for the test. This may improve the likelihood of clinical agreement. In addition, the larger boluses and the rapid, consecutive swallows required in the 3-oz. swallow test may produce more obvious results. Although reliable, the use of this measure has been questioned for putting patients at risk [5]. Further research is needed with this measure, and the risk involved needs to be weighed against the potential benefits.

An overall judgment of dysphagia was also made with significant inter- and intrajudge reliability. This complicates the issue of reliability because little can be determined regarding the origin of this judgment. Several measures may contribute to an overall rating of dysphagia even when some of the measures are individually unreliable. Perhaps there is strength in numbers, and the measures, collectively, produce a reliable, overall rating of dysphagia. If true, this challenges development of a clinical examination that is efficient and effective. Thus, further research needs to determine whether a few reli- able individual measures are sufficient to produce a re- liable rating of dysphagia.

Clinical examinations in this investigation were conducted according to reports of actual clinical practice [1], without pre-training to standard criteria. Our results indicate that clinicians can judge reliably fewer than 50% of the measures commonly employed in a clinical/bedside examination of swallowing. Oral motor and voice measures were rated more reliably than history or trial swallow measures. We suspect poor inter- and intrajudge reliability for some clinical measures in this investigation are the result of both patient and clinician variability. Clinician variability could be reduced through training. We know from studies of VFS measures that training influences reliability positively [42]. Such methods of training also should be examined for clinical measures. Low reliability in this study at the very least indicates a need to clearly define clinical measures and describe how those populations should be rated. In addition, future research involving clinical/bedside measures or swallowing should report inter- and intrajudge reliability for all measures. It has become standard practice to report reliability for VFS examinations of swallowing in research. Poor reliability in this investigation indicates that clinical measures of swallowing evaluation should be held to the same standard.

Acknowledgment. This work was funded in part by a grant from the Department of Veterans Affairs, Office of Academic Affiliations.

References

1. McCullogh GH, Wertz RT, Rosenbek JC, Dineen C: Clinicians' preferences and practices in conducting clinical/bedside and video fluoroscopic examinations of swallowing in an adult, neurogenic population. *Am J Speech Lang Pathol 8:*149–163, 1999
2. Linden P, Kuhlemeier KV, Patterson C: The probability of correctly predicting subglottic penetration from clinical observations. *Dysphagia 8:*170–179, 1993
3. Horner J, Brazer SR, Massey EW: Aspiration in bilateral stroke patients: a validation study. *Neurology 43:*430–433, 1993
4. Daniels SK, McAdam CP, Brailey K, Foundas AL: Clinical assessment of swallowing and prediction of dysphagia severity. *Am J Speech Lang Pathol 6:*17–24, 1997
5. Logemann JA, Veis S, Colangelo L: A screening procedure for oropharyngeal dysphagia. *Dysphagia 14:*44–51, 1999
6. Darley FL, Aronson AE, Brown JR: Differential diagnostic patterns of dysarthria. *J Speech Hear Res 12:*246–269, 1969
7. Zyski BJ, Weisiger BE: Identification of dysarthria types based on perceptual analysis. *J Speech Hear Res 34:*285–293, 1991
8. Sheard C, Adams RD, Davis PJ: Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *J Speech Hear Res 34:*285–293, 1991
9. Hammen VL, Yorkston KM, Dowden P: Index of contextual intelligibility: impact of semantic context in dysarthria. In: Moore C, Yorkston K, Beukelman D (eds.): *Dysarthria and Apraxia of Speech: Perspectives on Management.* Baltimore, MD: Paul H. Brooks, 1991, pp 65–75
10. Platt LJ, Andrews G, Young M, Quinn PT: Dysarthria of adult cerebral palsy: intelligibility and articulatory impairment. *J Speech Hear Res 23:*28–40, 1980
11. Tjaden K, Liss JM: The influence of familiarity on judgements of treated speech. *Am J Speech Lang Pathol 4:*39–47, 1995
12. Tjaden K, Liss JM: The role of listener familiarity in the perception of dysarthric speech. *Clin Ling Phonetics 9:*139–154, 1995
13. Blaustein S, Bar A: Reliability of perceptual voice assessment. *J Commun Disord 16:*157–161, 1983
14. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS: Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res 36:*21–40, 1993
15. Bassich CJ, Ludlow CL: The use of perceptual methods by new clinicians for assessing voice quality. *J Speech Hear Disord 51:*125–133, 1986
16. Gelfer MP: Perceptual attributes of voice: development and use of rating scales. *J Voice 2:*320–326, 1988
17. Krom G: Consistency and reliability of voice quality ratings for different types of speech fragments. *J Speech Hear Res 37:*985–1000, 1994
18. Logemann JA: *Evaluation and Treatment of Swallowing Disorders*, 2nd ed. Austin, TX: Pro-Ed, 1998
19. Langmore SE: Swallowing in an elderly veteran population. Presented at the 3rd Annual Dysphagia Research Society Meeting, McClean, VA, 1994
20. Splaingard ML, Hutchins B, Sulton LD, Chaudhuri G: Aspiration in rehabilitation patients: videofluoroscopic versus bedside clinical assessment. *Arch Phys Med Rehabil 69:*637–640, 1988
21. Croghan JE, Burke EM, Caplan S, Denman S: Pilot study of 12-month outcomes of nursing home patients with aspiration on videofluoroscopy. *Dysphagia 9:*141–146, 1994
22. Campbell-Taylor I, Fisher RH: The clinical case against tube feeding in palliative care of the elderly. *J Am Geriatr Soc 36:*1100–1104, 1987
23. Kohn CL, Keithley JK: Eneteral nutrition: potential complications and patient monitoring. *Nurs Clin North Am 24:*339–350, 1989
24. Cogen R, Weinryb J: Aspiration pneumonia in nursing home patients fed via gastrostomy tubes. *Am J Gastroenterol 84:*1509–1512, 1989
25. Buckwalter JA, Sasaki CT: Effect of tracheotomy on laryngeal function. *Otolaryngol Clin North Am 17:*41–48, 1984
26. DeVita MA, Spierer-Rundback L: Swallowing disorders in patients with prolonged orotracheal intubation or tracheostomy tubes. *Crit Care Med 18:*1328–1330, 1990
27. Elpern E, Jacobs E, Bone R: Incidence of aspiration in tracheally intubated adults. *Heart Lung 16:*527–531, 1987
28. Nash M: Swallowing problems in the tracheotomized patient. *Otolaryngol Clin North Am 21:*701–709, 1988
29. Jimenez P, Torres A, Rodriguez-Roisin R, Pugiz de la Bellacase JP, Aznar R, Gatell JM, Agusti-Vidal A: Incidence and etiology of pneumonia acquired during mechanical ventilation. *Crit Care Med 17:*882–885, 1989
30. Langmore SE, Terpenning MS, Schork A, Chen Y, Murray JT,

Lopatin D, Loesche WJ: Predictors of aspiration pneumonia: how important is dysphagia? *Dysphagia 13:*69–81, 1998

31. Feinberg MJ, Ekberg O, Segall L, Tully MA: Deglutition in elderly patients with dementia: findings of videofluorographic evaluation and impact on staging and management. *Radiology 183:*811–814, 1992
32. Issa FG: Gustatory stimulation of the oropharynx fails to induce swallowing in sleeping dogs. *Gastroenterology 107:*650–656, 1994
33. Vaughan GG, Grycko RJ, Montgomery MT: The prevention and treatment of aspiration and vomitus during pharmacosedation and general anesthesia. *J Oral Maxillofac Surg 50:*874–879, 1992
34. Stein M, Williams A, Grossman F, Weinberg A, Zuckerbraun L: Cricopharyngeal dysfunction in chronic obstructive pulmonary disease. *Chest 97:*347–352, 1990
35. Horner J, Massey EW, Brazer SR: Aspiration in bilateral stroke patients. *Neurology 40:*1686–1688, 1990
36. Leder SB: Videofluoroscopic evaluation of aspiration with visual examination of the gag reflex and velar movement. *Dysphagia 12:*1221–1223, 1997
37. Horner J, Massey EW, Riski JE, Lathrop DL, Chase KN: Aspiration following stroke: clinical correlates and outcome. *Neurology 38:*1359–1362, 1988
38. Gordon C, Hewer RL, Wade DT: Dysphagia in acute stroke. *BMJ 295:*411–414, 1987
39. Martin BJ, Corlew MM: The incidence of communication disorders in dysphagic patients. *J Speech Hear Disord 55:*28–32, 1990
40. Lof GL, Robbins JA: Test-retest variability in normal swallowing. *Dysphagia 4:*236–242, 1990
41. DePippo KL, Holas MA, Reding MJ: Validation of the 3-oz water swallow test for aspiration following stroke. *Arch Neurol 49:*1259–1261, 1992
42. Scott A, Perry A, Bench J: A study of interrater reliability when using videofluoroscopy as an assessment of swallowing. *Dysphagia 13:*223–227, 1998