



Published in final edited form as:

*Dysphagia*. 2009 June ; 24(2): 211–217. doi:10.1007/s00455-008-9196-6.

## Comparison of Trained Clinician Ratings with Expert Ratings of Aspiration on Videofluoroscopic Images from a Randomized Clinical Trial

Jacqueline A. Hind, M.S.<sup>1,2</sup>, Gary Gensler, M.S.<sup>3</sup>, Diane K. Brandt<sup>3</sup>, Patricia J. Miller Gardner, J.D.<sup>4</sup>, Loreen Blumenthal, M.S.P.A.<sup>5</sup>, Gary D. Gramigna, M.S.<sup>6</sup>, Steven Kosek, M.S.<sup>7</sup>, Donna Lundy, Ph.D.<sup>8</sup>, Susan McGarvey-Toler, M.S.<sup>9</sup>, Susan Rockafellow, M.S.<sup>10</sup>, Paula A. Sullivan, M.S.<sup>1,2</sup>, Marybell Villa, M.A.<sup>11</sup>, Gary D. Gill, Ph.D.<sup>1,2</sup>, Anne S. Lindblad, Ph.D.<sup>3</sup>, Jeri A. Logemann, Ph.D.<sup>12</sup>, and JoAnne Robbins, Ph.D.<sup>1,2</sup>

<sup>1</sup>University of Wisconsin, Madison, WI

<sup>2</sup>William S. Middleton Memorial Veterans Hospital, Madison, WI

<sup>3</sup>The EMMES Corporation, Rockville, MD

<sup>4</sup>American Speech- Language-Hearing Association, Rockville, MD

<sup>5</sup>West Palm Beach VA Medical Center, West Palm Beach, FL

<sup>6</sup>VA Boston Healthcare System, West Roxbury, MA

<sup>7</sup>Minneapolis VA Medical Center, Minneapolis, MN

<sup>8</sup>University of Miami Miller School of Medicine, Miami, FL

<sup>9</sup>Richard L. Roudebush VA Medical Center, Indianapolis, IN

<sup>10</sup>Westview Hospital, Indianapolis, IN

<sup>11</sup>New York Hospital Queens, Flushing, NY

<sup>12</sup>Northwestern University, Evanston, IL

### Abstract

Accurate detection and classification of aspiration is a critical component of videofluoroscopic swallowing evaluation, the most commonly utilized instrumental method for dysphagia diagnosis and treatment. Currently published literature indicates that inter-judge reliability for the identification of aspiration ranges from poor to fairly good depending on the amount of training provided to clinicians. The majority of extant studies compared judgments among clinicians. No studies included judgments made during the use of a postural compensatory strategy. The purpose of this study was to examine the accuracy of judgments made by speech-language pathologists (SLPs) practicing in hospitals compared with unblinded expert judges when identifying aspiration and using the 8-point Penetration/Aspiration Scale. Clinicians received extensive training for the detection of aspiration and minimal training on use of the Penetration/Aspiration Scale. Videofluoroscopic data were collected from 669 patients as part of a large, randomized clinical trial and include judgments of 10,200 swallows made by 76 clinicians from 44 hospitals in 11 states. Judgments were made on swallows during use of dysphagia compensatory strategies: chin down posture with thin-liquids and thickened liquids (nectar-thick and honey-thick consistencies) in a head neutral posture. The subject

population included patients with Parkinson's disease and/or dementia. Kappa statistics indicate high accuracy for all interventions by SLPs for identification of aspiration (all  $K > .86$ ) and variable accuracy (range 69%–76%) using the Penetration/Aspiration Scale when compared to expert judges. It is concluded that while the accuracy of identifying the presence of aspiration by SLPs is excellent, more extensive training and/or image enhancement is recommended for precise use of the Penetration/Aspiration Scale.

## Keywords

deglutition; aspiration; accuracy; clinical trial; Penetration/Aspiration Scale

---

## Introduction

Videofluoroscopic imaging of oropharyngeal swallowing is a widely utilized instrumental method for dysphagia diagnosis and treatment planning. Accurate detection and classification of aspiration is a critical component of this procedure. Inter-clinician reliability in the identification of aspiration using this radiographic procedure has been investigated by others<sup>1–5</sup>. A relatively small number of clinician judges participated in these studies with varying amounts or no training and few videofluoroscopic evaluations were judged. No studies were found that included more than 10 clinician judges, and with the exception of McCullough, et al., all used clinicians from the same facility thus limiting generalizability of results.

A review of the relevant literature indicates that interobserver reliability ranges from poor<sup>1–3</sup> to fairly good.<sup>4</sup> McCullough, et al., reported only “moderate” interobserver reliability in the identification of aspiration in one of 12 bolus conditions and either “slight” or “fair” reliability with the other 11 bolus conditions. Additionally, that study reported poor reliability of scores made by clinicians using the Penetration/Aspiration Scale (P/A Scale) to quantify the depth of airway invasion and patient response.<sup>6;7</sup> However, clinicians in that study were not pre-trained relative to the criterion of the judgments. Scott and colleagues studied the effect of training nine clinicians on the criteria used to identify aspiration along with other measures and reported a significant improvement in reliability with training.<sup>8</sup>

Kuhlmeier, et al., used both speech-language pathologists (SLP) and physicians to evaluate videofluoroscopic swallowing assessments and reported the greatest reliability for the detection of aspiration with thin liquids compared with other materials including nectar-thick liquid.<sup>3</sup> In all of these published studies, a variety of barium materials were used in published studies making comparisons difficult because of the lack of standardization.

The data presented here are from a large, NIH-funded multi-site clinical trial. Judgments made by trained SLPs working in acute care settings were compared with judgments made by expert judges at a central data collection laboratory. Judgments were made regarding the presence/absence of aspiration and also the anatomic depth of airway invasion and patient response using the P/A Scale.<sup>6;7</sup>

## Methods

Relevant details of the design and methods of this clinical trial are reported in detail elsewhere and are summarized here.<sup>9</sup> Inclusion criteria were a diagnosis of stable or progressive dementia (Alzheimer's type, single or multi-stroke type, or other non-resolving type) and/or Parkinson's disease, and an age range of 50 to 95 years. Participants were referred for a videofluoroscopic swallowing study (VFSS) by their SLP who, after completing a non-instrumental swallow screening, suspected aspiration with thin liquids. Bedside swallow symptoms warranting

referral for instrumental evaluation included, but were not limited to throat clearing, cough, choking, or wet/gurgly vocal quality during or after liquid swallows. Exclusion criteria included confounding medical issues including current tobacco or alcohol abuse, head or neck cancer, >20-year insulin-dependent diabetes, other progressive/infectious neurological disease, or pneumonia within the last six weeks. Patients were eligible to participate in this study if they were observed to aspirate thin liquids in the absence of any intervention during the VFSS. Aspiration was defined as barium observed below the vocal folds.

Once eligibility was determined, patients were given one to three boluses each of 3ml from a spoon and patient controlled amounts from a cup of standardized materials using the three discrete interventions in random order (Table 1):

1. Thin liquid (Varibar Thin®) with chin down posture (15 centipoise (cP))
2. Nectar-thick liquid (Varibar Nectar®), swallowed in head upright (neutral) position (300cP)
3. Honey-thick liquid, (Varibar Honey®), swallowed in head upright (neutral) position (3000cP)

### **Clinician Training on Judgment of Aspiration and Use of P/A Scale**

Prior to participation in the study, acute care clinicians completed two days of intensive training at the Central Laboratory (CL) at the University of Wisconsin-Madison during which they received instruction regarding the study design and data collection protocol including the criteria for judging the presence of aspiration. Each clinician successfully passed a test (>90% correct) for which they identified the presence or absence of aspiration on 25 videofluoroscopically-recorded swallows. Information regarding use of the P/A Scale was considerably less intense and consisted of 20 minutes of instruction about the scale score definitions and presentation of one videofluoroscopically recorded example of each score on the scale.

The videofluoroscopic data recorded either on S-VHS videotapes or digitally on a Kay Swallowing Workstation (KayPentax) were reviewed and judged by the treating SLP for the presence/absence of aspiration. In addition, clinicians were asked to rate each swallow using the P/A Scale (Table 2). The original tapes and digital images were then sent to the CL where each swallow was analyzed for the same measures within 24 hours. Judges in the CL were considered experts in using these measures as they completed rigorous training with one of the developers of the P/A Scale using training tapes and achieved inter- and intra-judge reliability of greater than 92%. The CL experts were not blinded to the clinician judgments because of the need to remain oriented as to which bolus was being judged using the clinician's form. If a clinician and CL expert disagreed about aspiration and it was considered by the expert to be a difficult swallow to judge, a second expert was asked to review the swallow and was blinded to both the clinician and first expert's judgments.

The equipment used in the CL was a Panasonic SVHS AG-7300 VCR that allowed for real time, slow motion and frame-by-frame analysis. Enhancement of images was at the discretion of the CL expert using an I.Den Digital Time Base Corrector IVT-9Plus and modified SVHS VCRs that allowed for contrast and intensity adjustment. Image quality did vary across the 44 sites; however, both the CL and the clinician judges viewed the same original images. Contrast enhancement was not performed on the images recorded for the 4% (n=27) of patients who used the Kay Digital Workstation due to instrumental constraints.

Seventy-six clinicians from 44 hospitals in 11 states collected data from 669 patients with a total data set of 10,200 swallows on which the binary presence/absence of aspiration was

judged. A subset of 50 clinicians also completed P/A Scale measurements for 369 subjects (4,191 swallows) in addition to the presence/absence of aspiration. Informed consent was obtained from all patients or from the patient's representative and each clinical facility had oversight by an Institutional Review Board (IRB) approved by the Office for Human Research Protections and the Department of Health and Human Services (DHHS).

### Statistical Methods

Kappa statistics were used to determine the degree of accuracy between the ratings of the pooled clinicians and the CL experts with regard to the presence of aspiration for each intervention (overall and for each swallow condition) and Kendall's tau correlations were used to compare the accuracy between specific P/A score, where data were available.<sup>10,11</sup> Chi-square statistics were used to compare accuracy across categories, including: age, gender, race, education, presence of gastrostomy, diagnosis, and order of administration of the swallowing interventions. There was no adjustment for multiple comparisons. All statistical analyses were done by using SAS software, version 8.2 (SAS Institute, Cary, North Carolina).

### Results

Demographic characteristics have been shown previously for this cohort.<sup>12</sup> Briefly, 70% percent were male, 59% were age 80 years and older, 15% were minorities, and 63% graduated from high school. Forty-nine percent had some form of dementia, 32% had Parkinson's disease without dementia, and 19% had Parkinson's disease with dementia.

Table 3 shows the agreement between the clinician's assessment and the expert's assessment of aspiration on each intervention (chin down posture with thin liquid, nectar-thick liquid or honey-thick liquid) both overall (i.e., on either swallow condition) and within each type of swallow condition (3ml or cup). Agreement was very high for each intervention overall (K =0.89 for chin down, K=0.89 for nectar-thick, K=0.86 for honey-thick) and high for both the 3ml swallows (K=0.79 for chin down, K=0.78 for nectar-thick, K=0.77 for honey-thick) and the cup swallows (K=0.85 for chin down, K=0.84 for nectar-thick, K=0.79 for honey-thick).

Table 4 shows the agreement between the SLP and CL expert assessment of aspiration on each swallow intervention across various categories (i.e. sex, age, race, order of administration). There was no difference in Kappa statistics between subject categories. An examination of categorical differences within each swallow condition (3ml and cup) reported similar results (not shown).

Table 5 shows the agreement between the clinician's and the CL expert's assessments of the P/A scores for the same swallow across interventions and conditions. Agreement is similarly high for each intervention regardless of swallow condition ( $\tau$  range: 0.73 to 0.81). Overall, exact agreement ranged from 69%-76% between the clinician and CL expert on the P/A scale score. Allowing for up to a one-step difference in the P/A improved the accuracy to 81%-86% across interventions and conditions. Within the chin down intervention, exact agreement was more likely to occur when the clinician gave a P/A score of aspiration (6-8). On the nectar-thick and honey-thick interventions, exact agreement was most likely to occur when the clinician initially gave a P/A score of no penetration (1) and least likely to occur when the clinician gave a P/A score of penetration (2-5), the disparity being far greater for the honey thick intervention. There were no categorical differences (age, gender, etc.) with respect to accuracy (data not shown).

## Discussion

High accuracy by trained SLPs for identification of aspiration and variable accuracy quantifying airway invasion using the P/A Scale was found on a series of videofluoroscopic swallowing evaluations as part of a large, multi-site clinical trial. While studies of interobserver reliability have been published with varying results – many indicating poor reliability - this study of the accuracy of clinicians in identifying and scoring aspiration and penetration compared with an expert is by far the most encouraging and the largest of its kind including 76 clinicians from 44 hospitals in 11 states judging over 10,000 swallows. This design was chosen so that the accuracy of scores relative to a standard could be quantified, rather than basing reliability on ratings between non-expert clinicians.

One recent study by Kelly et al. also reported encouragingly high interrater reliability (.79) compared P/A scores from videofluoroscopic images of two boluses for fifteen patients (30 total swallows) and rated by fifteen SLPs.<sup>5</sup> While the amount of pre-training regarding the P/A scale was not specified, the study did report the amount of SLP experience as 6.0 mean examination-years (one examination-year represents one examination performed and interpreted per week for one year – approximately 50 examinations), the P/A scores were reported as means and thus a detailed analysis of how reliability varied across the 8-point scale was not represented. Another key difference in this study, as with many others, is that non-expert SLP scores were compared with each other rather than with an expert indicating that while the judgments are consistent, determination of accuracy must be inferred.

The study herein was the only one found which included accuracy of judgments using a compensatory postural strategy. Results indicated that clinicians were as accurate identifying aspiration when patients were using the chin down posture as when the patient was in a head neutral position. [Move this next sentence?] As one of the primary purposes of the videofluoroscopic evaluation is assessing the effectiveness of compensatory strategies, it is reassuring now to be able to place confidence in the judgments of trained SLPs. Standardization of materials presented was achieved by using pre-packaged barium materials (E-Z-EM Varibar®) with known and consistent viscosities. While the present findings were limited to liquids, further examination of semi-solid and solid materials is warranted.

We found a high rate of agreement between the clinician and the CL experts for assessment of aspiration vs. no aspiration. Kappa statistics for overall aspiration were slightly higher than the kappa statistics for either swallow condition (3ml or cup), as the subject could aspirate on either type of swallow to be considered an aspirator on that intervention. The agreement was substantially high so as to make it unlikely that there could be differences among patient characteristics or order of bolus administration. Due to study design requiring the CL expert to evaluate the exact same swallows as the SLP within 24 hours of tape arrival to ensure participant safety, CL experts were not blinded to the judgments made by clinicians, which could also add to the high rate of accuracy. The CL experts had the ability instrumentally to enhance the videofluoroscopic images and improve contrast/brightness. Even in light of this ability by the CL to potentially make more accurate judgments, the agreement with clinicians continued to be excellent. It stands to reason that if the SLPs had the ability to enhance the images, their accuracy would be even higher.

P/A Scale Scores were determined about 41% of the time by clinicians. It was not a protocol requirement that they complete the measure, although they were encouraged to do so. While clinicians received minimal training in using the P/A scale, the level of accuracy was relatively high (tau correlations between 0.73 and 0.81) particularly at the high and low range of scores. Overall accuracy within one step using the P/A Scale was 81%–86%. It is unclear whether this

high level of accuracy was a true measure of clinician capability, or a function of clinician selectivity (i.e., potentially choosing to send scores for patients with easier swallows to judge).

Of note is that accuracy using the P/A Scale was considerably less for the mid-range of the scale in scores of four or five. Scores of four and five, indicating deep penetration, are relatively rare in that they accounted for no more than 6% of all scores. The infrequency of these two scores is consistent with what the original authors of the P/A Scale reported.<sup>6,7</sup> It could be that accuracy of judgments decreases when clinicians have less frequent opportunity to use the scores. Also, deep penetration is likely to occur with small amounts of barium so the experts' ability to enhance the contrast may have played a role in the discrepancies between clinician and expert.

On chin down interventions, exact agreement was more likely to occur when the clinician gave a P/A score of aspiration (6–8) (77%). A likely reason is that the trachea (where aspiration is scored) is comparatively more visible and less directly affected by the postural change as compared to the laryngeal vestibule (where penetration is scored), which may be more challenging. This would be consistent with prior studies that reported decreased distance between epiglottis and posterior pharyngeal wall with the chin down posture.<sup>13,14</sup> Measures of the anatomic movements relative to chin down posture and head neutral are forthcoming from this large, randomized clinical trial and hopefully will provide additional insight into these results.

Exact accuracy on the nectar and honey interventions was most likely to occur when the clinician gave a P/A score of no penetration (77% and 86% respectively) and least likely to occur when the clinician gave a P/A score of penetration (2–5) (69% and 61% respectively). Typically, the penetration scores by the SLP were 2 or 3, so this may be a function of not being able to differentiate between no penetration and minimal penetration.

The value of the P/A Scale is its ability to more discretely quantify swallowing than the typical gross categories of only the presence or absence of aspiration. Therefore, these findings suggest the need for additional clinician training and/or image enhancement when using the P/A Scale to achieve the reliable assignment of scores (particularly in the mid-range scores) and guide best dysphagia management.

Overall, this study supports the conclusion that trained speech-language pathologists can and do make more reliable judgments of aspiration than has been previously reported for untrained clinicians. Furthermore, since only 20 minutes of training yielded the encouraging result of 69%–76% overall accuracy using the P/A Scale, further investigation of more intensive training to raise the accuracy of P/A scores and improve dysphagia diagnosis, treatment and communication among practitioners is warranted. Additionally, these data provide critical quantitative evidence of professional quality to our colleagues, funding sources and third party payors which may prove crucial in the ever-evolving healthcare environment..

## Acknowledgments

Thank you to Megan Ballantine, Judy Bradferd, Jim Coyle, Karen Dikeman, Lorraine Downey, Meredith Gerdin, Barbara Grande, Jordyce Natzke, Jamie Pulliza, Dianne Starke, Jennifer Virnoche and Janine Weisberg for their significant contribution as clinician judges and to Abby Duane for manuscript preparation.

Supported by Grant # DC03206 to the American Speech-Language-Hearing Association, Communication Sciences and Disorders Clinical Trials Research Group from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Department of Health and Human Services, with additional support for the grant provided by Novartis and E-Z-EM, Inc. This material is the result of work supported with resources and use of facilities at the William S. Middleton Memorial Veterans Hospital, Madison, WI. This is GRECC Manuscript # 2007-16.



## Reference List

1. Wilcox F, Liss JM, Siegel GM. Interjudge agreement in videofluoroscopic studies of swallowing. *J Speech Hear Res* 1996;39:144–152. [PubMed: 8820706]
2. McCullough GH, Wertz RT, Rosenbek JC, Mills RH, Webb WG, Ross KB. Inter- and intrajudge reliability for videofluoroscopic swallowing evaluation measures. *Dysphagia* 2001;16:110–118. [PubMed: 11305220]
3. Kuhlemeier KV, Yates P, Palmer JB. Intra- and interrater variation in the evaluation of videofluorographic swallowing studies. *Dysphagia* 1998;13:142–147. [PubMed: 9633153]
4. Perlman AL, Booth BM, Grayhack JP. Videofluoroscopic predictors of aspiration in patients with oropharyngeal dysphagia. *Dysphagia* 1994;9:90–95. [PubMed: 8005013]
5. Kelly AM, Drinnan MJ, Leslie P. Assessing penetration and aspiration: How do videofluoroscopy and fiberoptic endoscopic evaluation of swallowing compare? *Laryngoscope* 2007;117:1723–1727. [PubMed: 17906496]
6. Rosenbek JC, Robbins JA, Roecker EB, Coyle JL, Wood JL. A Penetration-Aspiration Scale. *Dysphagia* 1996;11:93–98. [PubMed: 8721066]
7. Robbins J, Coyle J, Roecker E, Rosenbek J, Wood J. Differentiation of normal and abnormal airway protection during swallowing using the penetration-aspiration scale. *Dysphagia* 1999;14:228–232. [PubMed: 10467048]
8. Scott A, Perry A, Bench J. A study of interrater reliability when using videofluoroscopy as an assessment of swallowing. *Dysphagia* 1998;13:223–227. [PubMed: 9716754]
9. Brandt DK, Hind JA, Robbins JA, et al. Challenges in the design and conduct of a randomized study of two interventions for liquid aspiration. *Clinical Trials* 2006;3:457–468. [PubMed: 17060219]
10. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37–46.
11. Kendall M. A New Measure of Rank Correlation. *Biometrika* 1938;30:81–89.
12. Logemann JA, Gensler G, Robbins JA, et al. A randomized study of three interventions for aspiration of thin liquids in patients with dementia or Parkinson's disease. *JSLHR*. 2007 In Press.
13. Welch M, Logemann JA, Rademaker AW, Pauloski BR, Kahrilas PJ. Changes in pharyngeal dimensions affected by chin tuck. *Arch Phys Med Rehabil* 1993;74:178–181. [PubMed: 8431103]
14. Shanahan TK, Logemann JA, Rademaker AW, Pauloski BR, Kahrilas PJ. Chin-down posture effect on aspiration in dysphagic patients. *Arch Phys Med Rehabil* 1993;74:736–739. [PubMed: 8328896]

**Table 1**

## Interventions

	Head Posture	Presentation	
		3ml	Uncontrolled amount
Thin liquid (Varibar Thin®)	Chin-down	spoon	cup
Nectar-thick liquid (Varibar Nectar®)	Neutral	spoon	cup
Honey-thick liquid (Varibar Honey®)	Neutral	spoon	cup



Table 2

Penetration/Aspiration Scale<sup>6,7</sup>

Category	Score	Descriptions
<b>No Penetration or Aspiration</b>	<b>1</b>	<b>Contrast Does Not Enter the Airway</b>
<b>P E N E T R A T I O N</b>	<b>2</b>	<b>Contrast Enters the Airway, Remains Above Vocal Folds, No Residue</b>
	<b>3</b>	<b>Contrast Remains Above Vocal Folds, Visible Residue Remains</b>
	<b>4</b>	<b>Contrast Contacts Vocal Folds, No Residue</b>
	<b>5</b>	<b>Contrast Contacts Vocal Folds, Visible Residue Remains</b>
<b>A S P I R A T I O N</b>	<b>6</b>	<b>Contrast Passes Glottis, No Sub-Glottic Residue Visible</b>
	<b>7</b>	<b>Contrast Passes Glottis, Visible Sub-Glottic Residue Despite Patient's Response</b>
	<b>8</b>	<b>Contrast Passes Glottis, Visible Sub-Glottic Residue, Absent Patient Response</b>



**Table 4**

Comparison of agreement between clinician's assessment and the central laboratory's expert assessment of aspiration across categories, by intervention.

Category	Intervention		
	Chin Down	Nectar-thick	Honey-thick
Overall Kappa	0.89	0.89	0.86
Gender *			
Female	0.85	0.86	0.82
Male	0.91	0.90	0.88
Age *			
55–79	0.92	0.89	0.90
80–95	0.87	0.88	0.83
Race *			
White	0.90	0.89	0.86
Non-white	0.87	0.87	0.88
Education *			
< HS Graduate	0.85	0.90	0.84
HS Graduate	0.88	0.84	0.87
Some college	0.93	0.92	0.88
Gastrostomy *			
No	0.90	0.89	0.87
Yes	0.82	0.88	0.84
Diagnosis *			
Parkinson's disease	0.91	0.89	0.93
Dementia	0.87	0.90	0.81
Parkinson's disease with dementia	0.90	0.85	0.87
Order of administration *			
First	0.90	0.93	0.90
Second	0.87	0.89	0.88
Third	0.90	0.83	0.80

\* There are no statistically significant associations (p-value > 0.05 for all).

**Table 5**

Comparison of agreement between clinician’s assessment and the central laboratory’s expert assessment of Penetration-Aspiration Scale scores by intervention.

Intervention, condition	Clinician Assessment	Patients	Exact Match		Level of agreement between clinician and central laboratory assessments of pen-asp score <sup>1</sup>		Up to One Step Difference <sup>2</sup>
			N	%	N	%	
Chin-down, 3ml	No penetration (1)	250	168	67	227	91	
	Penetration (2–5)	248	155	63	207	83	
	Aspiration (6–8)	231	178	77	194	84	
	Overall	729	501	69	628	86	
	Tau correlation		0.77				
Nectar, 3ml	No penetration	336	255	76	303	90	
	Penetration	217	132	61	178	82	
	Aspiration	182	125	69	139	76	
	Overall	735	512	70	620	84	
	Tau correlation		0.74				
Honey, 3ml	No penetration	422	336	80	367	87	
	Penetration	166	97	58	130	78	
	Aspiration	181	132	73	138	76	
	Overall	769	565	73	635	83	
	Tau correlation		0.73				
Chin down, cup	No penetration	161	110	68	149	93	
	Penetration	238	162	68	194	82	
	Aspiration	227	175	77	197	87	
	Overall	626	447	71	540	86	
	Tau correlation		0.81				
Nectar, cup	No penetration	180	139	77	158	88	
	Penetration	237	157	66	188	79	

Intervention, condition	Clinician Assessment	Patients	Exact Match		Level of agreement between clinician and central laboratory assessments of pen-asp score <sup>1</sup>		Up to One Step Difference <sup>2</sup>	%
			N	%	N	%		
Honey, cup	Aspiration	207	145	70	161	78		
	Overall	624	441	71	507	81		
	Tau correlation		0.76					
	No penetration	344	297	86	316	92		
	Penetration	163	97	60	126	77		
	Aspiration	201	152	76	161	80		
Overall	708	546	76	603	85			
Tau correlation			0.77					

<sup>1</sup> Agreement is based on the exact P/A scale value assigned by the clinician and the CL. E.g., a score of 3 by the clinician and 3 by the CL indicates exact agreement, whereas a score of 2 by the clinician and 3 by the CL indicates a difference of one on the P/A scale.

<sup>2</sup> Includes all instances of exact match and all instances where the difference was exactly one step on the P/A scale