Inter-rater Reliability of the Dysphagia Outcome and Severity Scale (DOSS): Effects of Clinical Experience, Audio-Recording and Training.

Zarkada A[1,2], Regan J[3].

1Department of Clinical Speech and Language Studies, Trinity College Dublin, Dublin, Ireland. aggelikazark@gmail.com.

2Ziria, Aigialeia, Achaia, 25100, Peloponnese, Greece. aggelikazark@gmail.com.

Keywords: Deglutition; scale; inter-rater reliability; deglutition disorders

.

ABSTRACT

The Dysphagia Outcome and Severity Scale (DOSS) is widely used to measure dysphagia severity based on videofluoroscopy (VFSS). This study investigated inter-rater reliability (IRR) of the DOSS. It also determined the effect of clinical experience, VFSS audio-recording and training on DOSS IRR. A quantitative prospective research design was used. Seventeen speech and language pathologists (SLPs) were recruited from an acute teaching hospital, Dublin (> 3 years' VFSS experience, n = 10) and from a postgraduate dysphagia programme in a university setting (< 3 years' VFSS experience; n = 7). During testing, participants viewed eight VFSS clips (5 with audio-recording). Each VFSS clip was independently rated using the DOSS scale. Four weeks later, the less experienced group attended a 1-h training session on DOSS rating after which DOSS IRR was re-tested. Cohen's kappa co-efficient was used to establish IRR. IRR of the DOSS presented only fair agreement ($\kappa = 0.36$, $p < 0.05$). DOSS IRR was significantly higher ($\kappa = 0.342$) within the more experienced SLP group, compared to the less experienced SLP group ($\kappa = 0.298$) ($p < 0.05$). DOSS IRR was significantly higher in VFSS clips with audio-recording ($\kappa = 0.287$) compared to VFSS clips without audio-recording ($\kappa = -0.0395$) ($p < 0.05$). IRR of the DOSS pre-training ($\kappa = 0.328$) was significantly better comparing to post-training ($\kappa = 0.218$) ($p < 0.05$). Findings raise concerns as the DOSS is frequently used in clinical practice to capture dysphagia severity and to monitor changes.

INTRODUCTION

The Dysphagia Outcome and Severity Scale (DOSS) is a 7- point rating scale which measures dysphagia severity based on videofluoroscopy (VFSS) and makes recommendations for nutrition level, diet and independence (see Table 1). A single research study by O'Neil et al (1999) [1] has been conducted to date to evaluate DOSS inter-rater reliability (IRR). While it demonstrated that the DOSS presents high inter- and intra-rater reliability among speech and language pathologists (SLPs), this study was based on written reports of the VFSS procedure only. In clinical practice, marked uncertainty is reported amongst SLPs regarding scoring of the DOSS scale. Lack of DOSS rating agreement amongst SLPs is of great concern as these ratings impact on oral and alternative feeding decisions and on candidacy for dysphagia rehabilitation. Adequate reliability of the DOSS is therefore imperative so that SLPs can use the DOSS to accurately quantify dysphagia severity and guide clinical decision making in dysphagia practice. This challenge, coupled with an increased emphasis on the interpretation of the VFSS, set a need of obtaining further information regarding the IRR of the DOSS, exploring if clinical experience and training can improve the IRR of the DOSS. Moreover, it is unknown if the audio-recording in VFSS clips affects the IRR of the DOSS.

The primary purpose of this research was to establish the IRR of the DOSS based on VFSS recordings. The secondary purpose was to determine the impact of clinical experience, training and VFSS audio recording on DOSS IRR. The influence of these parameters on the reliability of different scales has been already investigated. The application of Penetration Aspiration Scale (PAS) for FEES demonstrates excellent inter- and intra-rater reliability regardless of clinician experience [2], while the Berg Balance Scale (BBS) illustrates high inter- and intra-rater reliability, even if raters have various levels of clinical experience [3]. As a consequence, research findings demonstrate that clinical experience may not affect the reliability. Different studies have shown that training has improved reliability of other scales in many clinical settings. An effective method to standardise the use of rating scales is the clinicians' training via video and certification [4]. Following this procedure, National Institutes of Health Stroke

Scale (NIHSS) presented moderate to excellent inter-rater agreement on most Stroke Scale items [4]. Similarly, the development of a video-based training package, including technical issues, patient selection procedures, and strategies of scoring and assessment had as a result the improvement of reliability of the Modified Rankin Scale (MRS) grading [5]. The Web-based feedback training program improved the reliability in clinicians' ratings of the Global Assessment of Functioning (GAF) Scale [6]. This finding concerns the clinicians in mental health practice who do not have a masters or doctoral degree. Observer Rating of Medication Taking (ORMT) scale can present satisfactory IRR in inpatient settings, given that raters have undertaken independent online training on this scale [7]. The MBSImp was characterized by high inter- and intra-rater reliability following standardized training for SLPs [8]. Consequently, existing studies demonstrate that training can improve the IRR with many ways. However, the most appropriate training method and time have not been established yet. According to Royal College of Speech and Language Therapists (RCSLT) (2007), recording equipment that incorporates audio recording, along with visual images constitutes an important standard of the methodology for conducting the VFSS procedure [9]. As a consequence, audio recording can ensure a better interpretation of VFSS clips, leading to better clinician's agreement. In clinical practice, it has been noticed that there is uncertainty regarding how the DOSS scale is scored.

The impact of SLPs' clinical experience, training and audio-recording of VFSS clips on IRR of the DOSS has not been investigated to date. As a consequence, the most appropriate conditions in which DOSS can demonstrate high IRR are currently unknown. This has an impact on patient care as recommendations for nutrition level, diet and independence are being made based on DOSS scores. Reliability of the DOSS is, therefore, imperative so that SLPs are able to use the DOSS to accurately quantify dysphagia severity and make suitable recommendations for nutrition, diet and independence. The present study addresses the gaps in the literature to date. The rationale for this study is to establish the IRR of the DOSS, to determine the impact of clinical experience on IRR of the DOSS, to determine the impact of training on IRR of the

DOSS and to determine the impact of audio-recording in VFSS clips on IRR of the DOSS. Authors hypothesize that clinical experience improves IRR of the DOSS and that audio-recording during VFSS impacts on DOSS IRR. It is also hypothesized that DOSS IRR can be improved with training.

METHODOLOGY

This study obtained ethical approval from the School of Linguistic, Speech and Communication Sciences Research Ethics Committee, Trinity College Dublin (TCD), in December of 2015.

Participants

The target population of this research was SLPs divided into two groups. The inclusion criteria of the first group were (i) certified training in VFSS analysis, (ii) a minimum three years' experience of working with dysphagia in an adult population. The inclusion criteria of the second group were (i) short training in VFSS analysis, (ii) limited experience (less than three years) of working with dysphagia in adult population.–The first group included ten more clinically experienced SLPs who were employees in an Acute Teaching Hospital, Dublin, Ireland, while the second group consisted of seven less clinically experienced SLPs who were M.Sc. and postgraduate diploma Dysphagia students, in the Department of Clinical Speech and Language Studies (CSLS), TCD, Ireland..

Protocol

The more clinically experienced SLPs were seated in a quiet room in the Speech and Language Therapy Department in the Acute Teaching Hospital, Dublin, Ireland. In another session, the less clinically experienced SLPs were seated in a quiet classroom in the Department of CSLS, TCD. Both of these SLP groups followed the exact same procedure. After completing the demographic section of a data collection form, participants were asked to watch eight VFSS clips (5 of which had audio recording), which were projected onto a large screen. The VFSS recordings were viewed both in real time (2 times) and in slow motion (1 time). Participants

could hear the audio-recording during VFSS clips via a speaker which was set at maximum for both groups. For each VFSS clip, study participants were provided with information regarding the patients' age and medical diagnosis (Myasthenia Gravis, Parkinson's disease, Myopathy, History of Reflux and Pneumonia, Respiratory Disease, No Medical Diagnosis, Stroke, Chronic Obstructive Pulmonary Disease (COPD) and Parkinsonism). VFSS clips included a variety of compensatory strategies as well as almost all consistencies of food (thin, medium, thick, puree, solid) as per patient's ability to swallow. VFSS clips corresponded with each level of the DOSS, and they presented adult's swallowing process.

Participants were asked to rate each VFSS clip on the DOSS scale without any conferring with colleagues. In order for DOSS rating of VFSS clips to be collated, both groups used "clickers" to attribute the severity level of dysphagia to each VFSS clip. As a consequence, participants rated each VFSS clip on the DOSS from 1 to 7, by selecting the appropriate button. When all participants had been collated by the system, they could see the group responses for each VFSS clip on the projector screen, that is to say the percentage of participants who had selected each dysphagia severity level to each VFSS clip. Each data collection session is estimated that it lasted approximately 1 hour and 45 minutes. At the end of the session, participants submitted completed data collection forms.

In order to investigate the effect of training on IRR of the DOSS, the less clinically experienced SLP group was asked to return for a one hour group training session. This was held four weeks later to avoid recollection of initial scoring. Training was provided by the researcher and the research supervisor. During training, participants observed, rated and discussed separate VFSS clips using the DOSS in an interactive group setting.

Upon completion of the one hour training session, the less clinically experienced SLP group was asked to rate the same eight VFSS clips as in the first session four weeks previously. Clips were presented in random sequence. Participants rated each VFSS clip using the DOSS scale, without any conferring with colleagues.

Statistical Analysis

In order for IRR of the DOSS to be established, the Cohen's kappa co-efficient was used. The level of significance was set at $p < 0.05$.All analyses were performed using R project.

RESULTS

Participant completion of VFSS is illustrated in Table 2. Results of DOSS ratings are presented in Table 3 and 4. Table 5 illustrates the various degrees of IRR established. DOSS Mean Rate from each rater category is presented in Figure 1.

It is noted that, during the data collection in the Acute Teaching Hospital, one of the raters discontinued after rating five of eight VFSS clips (three VFSS clips were not rated). Another rater in the Acute Teaching Hospital setting withdraw after rating seven of eight VFSS clips. In the first data collection in the Department of CSLS, TCD, one rater returned to rate the last three VFSS clips in a separate session two days later. In the training session and the second data collection, two raters withdrawn, remaining five participants.

From the seventeen SLPs who participated in the study ten of them were working in an Acute Teaching Hospital and seven of them were studying in the Department of CSLS, TCD. All participants were female.

From the ten SLPs who were working in an Acute Teaching Hospital, four of them were between 21 and 30 years old and six of them were between 31 and 40 years old. As for their educational level, six of them had Bachelor degree, three of them had M.Sc. and one of them had Ph.D. As for their specialised field, eight of the participants were working with the field of dysphagia and communication disorders and two of them were working with the field of dysphagia only. Furthermore, eight of the participants were working with adult population while two of them were working with both adult and pediatric population. Also, this SLP group had clinical experience in the field of dysphagia ranging from 3 years and above. This SLP

group had completed certified training in VFSS analysis and they were using the DOSS in their clinical placement.

From the seven SLPs who were studying in the Department of CSLS, TCD, five of them were between 21 and 30 years old and two of them were between 31 and 40 years old. As for their educational level, three of them had postgraduate diploma and four of them had M.Sc. As for their specialised field, one of the participants were working with the field of dysphagia and communication disorders and six of them were working with the field of dysphagia only. Furthermore, four of the participants were working with adult population while three of them were working with both adult and pediatric population. Also, six of the less clinically experienced SLPs had clinical experience in the field of dysphagia ranging from 1 to 3 years , while only one participant had not any clinical experience in the field of dysphagia. In addition, this SLP group had completed only a short training session in VFSS analysis, while only four participants were using the DOSS in their clinical placement.

Across all participants (n=17), the IRR of the DOSS was 0.36 (z = 19.9, p =0.001), presenting only fair agreement. IRR of DOSS ratings was significantly higher (κ = 0.342, z = 12.3, p = 0.001) within the more clinically experienced (Acute Teaching Hospital) SLP group, compared to the less clinically experienced (postgraduate) SLP group (κ = 0.298, z = 7.05, p =0.001). However, IRR of the DOSS is presented as fair in both cases. A one hour training session did not improve the IRR of the DOSS in the less clinically experienced SLP group. In fact, IRR before training (κ = 0.328, z = 6.37, p = 0.001) was significantly better comparing to post training (κ = 0.218, z = 4.22, p = 0.001). Despite this, IRR of the DOSS is presented as fair in both cases. IRR of the DOSS was significantly better using VFSS clips with audio-recording (κ = 0.287, z = 8.99, p = 0.001) compared to clips without audio (κ = -0.0395, z = -0.817, p = 0.414). As a consequence, in case that the VFSS clips had audio-recording, IRR of the DOSS is presented as fair, while in case that VFSS clips had not audio-recording, IRR of the DOSS is presented as less than chance.

DISCUSSION

The purpose of the present study is to investigate the IRR of the DOSS. It also seeks to research the parameters which impact on IRR of the DOSS, so that SLPs can rate VFSS clips more accurately, improving the consistency of documentation and recommendations across clinicians. The impact of clinical experience, training and VFSS audio recording on IRR of the DOSS was studied. The present results indicate that the DOSS presents only fair agreement across SLPs. This research suggests that SLTs' clinical experience and VFSS audio recording can play a significant role in the increase of IRR of the DOSS while the one hour training session may have negative outcome in the IRR of the DOSS.

In contrast with the initial research by O'Neil et al. (1999) that demonstrated high IRR (90%) of the DOSS [1], this study demonstrated fair agreement in total ($\kappa = 0.36$). The O'Neil et al. (1999) study [1] demonstrated the IRR of the DOSS based on the interpretation of written reports from a review of the VFSS. Instead, in the current study, participants should rate eight different VFSS clips while they were only aware of the patients' age and medical diagnosis that each VFSS clip illustrated. In O'Neil et al. (1999) study [1], the training of SLPs in DOSS scale was based on specific instruction in the guidelines for use of the DOSS. In contrast, in this study, only the SLPs with limited experience in VFSS analyses were trained in the DOSS scale, with the intention of exploring if training can improve the IRR of the DOSS amongst less clinically experienced SLPs. In this case, training was based on the attention of severity level headings and the analysis of original VFSS clips.

The present findings have clinical implications for SLPs who work with adults with dysphagia and they need a valid clinical tool in order to properly assess the dysphagia severity level, during the VFSS procedure. Moreover, the use of a reliable clinical tool may have an impact on clinicians about how to better diagnose and improve their communication for better care for the patients with dysphagia. In this study, the demonstration of fair IRR of the DOSS illustrates that SLPs rate the severity of dysphagia with different way and as a result possible changes in

patients' swallow function (spontaneous recovery, response to rehabilitation or deterioration) may not be identified consistently across SLPs. So, the safest and most beneficial care that should be provided to the patients are not be ensured. However, it is observed that the standard deviation is actually quite low for almost all VFSS clips, while raters of both groups selected contiguous ratings in most cases. As a consequence, we can come to the conclusion that even if participants rated the VFSS clips with different ratings, they are able to detect the clinical characteristics of dysphagia in each VFSS clip almost consistently. Moreover, it is noticed that raters present a larger variation of ratings in clips in which patients present Parkinson's disease, myopathy, history of reflux and pneumonia, no medical diagnosis and stroke as opposed to patients who present myasthenia Gravis, respiratory disease as well as Chronic Obstructive Pulmonary Disease (COPD) and parkinsonism. This fact may lead to the conclusion that the consistency across clinicians depends on the diagnosis of each patient.

According to O' Neil et al. (1999) [1], the DOSS rating scale is the only scale that describes the severity dysphagia level, taking into consideration the three aspects of swallowing making recommendations for nutrition, diet, and independence. Currently existing scales which have been based on too general and subjective definitions per level have not included all important dysphagia issues or have not presented adequate degree of reliability [1]. Furthermore, the present findings focused on the effect of SLPs' clinical experience, VFSS audio-recording and a short training session on IRR of DOSS ratings.

IRR of DOSS ratings was significantly higher within the more clinically experienced SLP group, compared to the less clinically experienced SLP group. These findings come in contrast with other studies that illustrated that rating scales demonstrate high inter- and intra-rater reliability regardless of clinician experience [2, 3].

The one hour training session did not improve the IRR of the DOSS. In fact, IRR pre-training was significantly better comparing to post-training. The fact that training not only had not statistically significant impact in the demonstration of higher IRR of the DOSS but also

deteriorated the IRR of this scale comes in contrast with previous studies. These studies illustrated that training played an essential role in the increase of IRR of various scales [4- 8]. However, no one study illustrates that one hour is enough time for an efficient training.

Also, IRR of the DOSS was significantly better in VFSS clips with audio-recording compared to clips without audio-recording. Many clinical settings do not have audio-recording during VFSS procedure and this could impact on IRR. This could be assured by the RCSLT (2007) [9] view that VFSS that incorporates audio recording in addition to visual images constitutes an important standard of the methodology for conducting the VFSS procedure.

The findings of this research may therefore be used to demonstrate that the IRR of the DOSS is fair. As a consequence, more clinical experience and sufficient training may improve the IRR of this scale. Moreover, this research may highlight the necessity of audio-recording during the VFSS procedure.

Despite the effort expended to thoroughly study the issues that this research set as a purpose, there were some problems and limitations that could not be overcome. One of the main limitations of this research is the small sample size recruited and the small number of VFSS clips reviewed. Even if it provides a general image about the degree of IRR of the DOSS, it is impossible to generalize the results. Also, four out of seven less clinically experienced SLTs were using the DOSS in their clinical placement, while the remaining participants were not using the DOSS. The fact that only a number of participants were familiar with the DOSS can affect the results. As for more clinically experienced participants, all of them were certified trained in VFSS analysis and they were using the DOSS. Moreover, a number of participants withdrew during the session. More specifically, during the data collection in the Acute Teaching Hospital, one of the participants did not rate the last three VFSS clips and one of them did not rate the last one. As a consequence, the maximum number of unrated VFSS clips was three. In addition, during the data collection I in the Department of CSLS, TCD, participants rated all VFSS clips, but in the training session and data collection II, two of them withdrew.

That is the total agreement between all raters cannot be calculated accurately. The length of the training session in the DOSS rating scale was limited to one hour. As a result, the potential effects of a longer and more in-depth training programme were not uncovered within this research.

Further studies are still needed with larger group of participants and larger number of VFSS clips, a more expanded interactive training session – at least 4 hours, and a possible modification of the DOSS form may improve inter-rater reliability of the DOSS. Finally, the investigation of the intra-rater reliability of the DOSS would be useful in clinical practice.

Declaration of interest statement: The authors declare that there is no conflict of interest

REFERENCES

1. O'Neil KH, Purdy M, Falk J, & Gallo L (1999) The dysphagia outcome and severity scale. Dysphagia 14(3), 139-145.
2. Butler SG, Markley L, Sanders B, Stuart A (2015) Reliability of the penetration aspiration scale with flexible endoscopic evaluation of swallowing. Annals of Otology Rhinology & Laryngology 0003489414566267.
3. Wong, C K (2014) Interrater reliability of the Berg Balance Scale when used by clinicians of various experience levels to assess people with lower limb amputations Physical therapy 94(3), 371-378.
4. Meyer B C, Lyden P D (2009) The modified National Institutes of Health Stroke Scale: its time has come. International Journal of Stroke 4(4), 267-273.
5. Quinn T J, Lees KR, Hardemark HG, Dawson J, Walters M R (2007) Initial experience of a digital training resource for modified Rankin scale assessment in clinical trials. Stroke 38(8), 2257-2261.
6. Støre-Valen, J, Ryum T, Pedersen G A, Pripp A H, Jose P E, Karterud, S (2015) Does a web-based feedback training program result in improved reliability in clinicians' ratings of the Global Assessment of Functioning (GAF) Scale? Psychological assessment, 27(3), 865.
7. Byrne M K, Deane F P, Murugesan, G, Connaughton E (2014) Interrater reliability of the Observer Rating of Medication Taking scale in an inpatient mental health facility. International journal of mental health nursing 23(6), 498-505.
8. Martin-Harris B, Brodsky M B, Michel Y, Castell DO, Schleicher M, Sandidge J, Blair J (2008) MBS measurement tool for swallow impairment—MBSImp: establishing a standard. Dysphagia 23(4), 392-405.

9. Royal College of Speech and Language Therapists (RCSLT) (2007) Videofluoroscopic Evaluation of Oropharyngeal Swallowing Disorders (VFS) in Adults: The role of Speech and Language Therapists. RCSLT Policy Statement. Retrieved from https://www.rcslt.org/docs/freepub/VFS_policy_statement_January_2007.pdf

Table 1: The Dysphagia Outcome and Severity Scale (DOSS) [1]

Full per-oral nutrition (P.O): Normal diet	**Level 7: Normal in all situations** • Normal diet • No strategies or extra time needed
	Level 6: Within functional limits/modified independence • Normal diet, functional swallow • Patient may have mild oral or pharyngeal delay, retention or trace epiglottal undercoating but independently and spontaneously compensates/clears • May need extra time for meal • Have no aspiration or penetration across consistencies
Full P.O: Modified diet and/or independence	**Level 5: Mild dysphagia**: Distant supervision may need one diet consistency restricted. May exhibit one or more of the following: • Aspiration of thin liquids only but with strong reflexive cough to clear completely • Airway penetration midway to cords with one or more consistency or to cords with one consistency but clears spontaneously • Retention in pharynx that is cleared spontaneously • Mild oral dysphagia with reduced mastication and/or oral retention that is cleared spontaneously

	Level 4: Mild–moderate dysphagia: Intermittent supervision/cueing, one or two consistencies restricted. May exhibit one or more of the following:
	• Retention in pharynx cleared with cue
	• Retention in the oral cavity that is cleared with cue
	• Aspiration with one consistency, with weak or no reflexive cough
	  o Or airway penetration to the level of the vocal cords with cough with two consistencies
	  o Or airway penetration to the level of the vocal cords without cough with one consistency

	Level 3: Moderate dysphagia: Total assist, supervision, or strategies, two or more diet consistencies restricted. May exhibit one or more of the following:
	• Moderate retention in pharynx, cleared with cue
	• Moderate retention in oral cavity, cleared with cue
	• Airway penetration to the level of the vocal cords without cough with two or more consistencies
	  o Or aspiration with two consistencies, with weak or no reflexive cough
	  o Or aspiration with one consistency, no cough and airway penetration to cords with one, no cough

Non-oral nutrition necessary	**Level 2: Moderately severe dysphagia**: Maximum assistance or use of strategies with partial P.O. only (tolerates at least one consistency safely with total use of strategies) May exhibit one or more of the following: • Severe retention in pharynx, unable to clear or needs multiple cues • Severe oral stage bolus loss or retention, unable to clear or needs multiple cues

	• Aspiration with two or more consistencies, no reflexive cough, weak volitional cough o Or aspiration with one or more consistency, no cough and airway penetration to cords with one or more consistency, no cough
	Level 1: Severe dysphagia: NPO: Unable to tolerate any P.O. safely May exhibit one or more of the following: • Severe retention in pharynx, unable to clear • Severe oral stage bolus loss or retention, unable to clear • Silent aspiration with two or more consistencies, nonfunctional volitional cough o Or unable to achieve swallow

Table 2. Participant completion of VFSS clips

VFSS clips	I	II	III	IV	V	VI	VII	VIII
Acute Teaching Hospital (*n*=10)	100%	100%	100%	100%	100%	90%	90%	80%
TCD Session I (*n*=7)	100%	100%	100%	100%	100%	100%	100%	100%
TCD Training Session (*n*=7)	70%	70%	70%	70%	70%	70%	70%	70%
TCD Session II (*n*=7)	70%	70%	70%	70%	70%	70%	70%	70%

Table 3. DOSS Ratings, Mean Rating, Standard Deviation and Agreement amongst SLPs (*n*=17) in the DOSS Rating of each VFSS Clip

VFSS clip	**DOSS Ratings**							***μ***	***St.D.***	**Agreement**
	1	**2**	**3**	**4**	**5**	**6**	**7**			
I	-	11	6	-	-	-	-	2.4	0.50	51.5%
II	-	-	-	1	11	5	-	5.27	0.59	47.8%
III	2	8	6	1	-	-	-	2.33	0.81	32.6%
IV	-	-	-	4	13	-	-	4.73	0.45	61%
V	-	-	-	-	-	10	7	6.40	0.50	48.5%
VI	-	-	1	8	7	-	-	4.33	0.61	36%
VII	6	9	-	1	-	-	-	1.73	0.79	37.5%
VIII	-	-	8	7	-	-	-	3.47	0.51	36%
	Total agreement									43.8%

μ, Mean; St.D., Standard Deviation

Table 4. DOSS Ratings, Mean Rating, Standard Deviation and Agreement amongst each rater category in the DOSS Rating of each VFSS Clip

More Clinically Experienced SLPs (n=10)										
VFSS Clip	**DOSS ratings**							***μ***	***St.D.***	**Agreement**
	1	**2**	**3**	**4**	**5**	**6**	**7**			
I	-	7	3	-	-	-	-	2.38	0.51	49.1%
II	-	-	-	-	8	2	-	5.25	0.46	56.3%
III	2	4	3	1	-	-	-	2.25	1.03	25.4%
IV	-	-	-	3	7	-	-	4.63	0.51	56.3%
V	-	-	-	-	-	4	6	6.63	0.51	45.4%
VI	-	-	-	4	5	-	-	4.50	0.53	36.3%
VII	3	6	-	-	-	-	-	1.63	0.51	38.1%
VIII	-	-	4	4	-	-	-	3.50	0.53	29.1%
	Total agreement									42.1%
Less Clinically Experienced SLPs (n=7)										
VFSS Clip	**DOSS ratings**							**μ**	**St.D.**	**Agreement**
	1	**2**	**3**	**4**	**5**	**6**	**7**			
I	-	**4**	**3**	-	-	-	-	2.43	0.53	42.8%
II	-	-	-	**1**	**3**	**3**	-	5.29	0.75	28.5%
III	-	**4**	**3**	-	-	-	-	2.43	0.53	42.8%
IV	-	-	-	**1**	**6**	-	-	4.86	0.37	71.4%
V	-	-	-	-	-	**6**	**1**	6.14	0.37	71.4%
VI	-	-	**1**	**4**	**2**	-	-	4.14	0.69	33.3%
VII	**3**	**3**	-	**1**	-	-	-	1.86	1.06	28.5%
VIII	-	-	**4**	**3**	-	-	-	3.43	0.53	42.8%
	Total agreement									45.2%
Less Clinically Experienced SLPs (n=7); Post-training										
	DOSS ratings							**μ**	**St.D.**	**Agreement**

VFSS Clip	1	2	3	4	5	6	7			
I	-	**2**	**3**	-	-	-	-	2.60	0.54	40%
II	-	-	-	**1**	**2**	**2**	-	5.20	0.83	20%
III	-	**2**	**1**	**2**	-	-	-	3.00	1.00	20%
IV	-	-	-	**1**	**3**	**1**	-	5.00	0.70	30%
V	-	-	-	-	-	**2**	**3**	6.60	0.54	40%
VI	-	-	-	-	**3**	**2**	-	5.40	0.54	40%
VII	**1**	**4**	-	-	-	-	-	1.80	0.44	50%
VIII	-	-	-	**2**	**3**	-	-	4.60	0.54	40%
	Total agreement									35%

μ, Mean; St.D., Standard Deviation

Table 5. Cohen's kappa coefficient for each rater category and corresponding significance

Case		**Subjects**	**Raters**	***Kappa***	***Z***	***p-value***
Total		5[a]	17	0.36	19.9	0.001*
Degree of Clinical Experience	**More Clinically Experienced SLPs**	5	10	0.342	12.3	0.001*
	Less Clinically Experienced SLPs	8	7	0.298	7.05	0.001*
Training	**Before**	8	5[b]	0.328	6.37	0.001*
	After	8	5	0.218	4.22	0.001*
Audio-recording	**Audio**	3	17	0.287	8.99	0.001*
	No Audio	2	17	-0.0395	-0.817	0.414

a The number of subjects is less than 8 because some SLPs did not rate 3 of the 8 VFSS clips

b The total Less Clinically Experienced SLPs was less than 7 because 2 SLPs could not complete the research

*p <0.05

Figure 1: DOSS Mean Rate from each rater category