University of Nevada, Reno

Speech-Language Pathologists Ratings of the Yale 3-Ounce Water Swallow Challenge: Accuracy, Reliability, and Clinician Demographics

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in
Speech Pathology

by

Amanda L. Morrissey, Ph.D.(c)

Kristine Galek, Ph.D./Dissertation Advisor

December 2022

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

AMANDA L MORRISSEY

entitled

**Speech-Language Pathologists Ratings of the Yale 3-Ounce Water Swallow Challenge:
Accuracy, Reliability, and Clinician Demographics**

be accepted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

Kristine Galek, Ph.D.
Advisor

Thomas Watterson, Ph.D.
Committee Member

Tami Brancamp, Ph.D.
Committee Member

Li-Ting Chen, Ph.D.
Committee Member

Julie Smith-Gagen, Ph.D.
Graduate School Representative

Markus Kemmelmeier, Ph.D., Dean
Graduate School

December, 2022

Abstract

Objective: Examine Speech-Language Pathologists (SLP) water swallow challenge (WSC) ratings across different clinical scenario videos (CSVs). Method: A non-experimental cross-sectional correlational design included eight expert and 150 non-expert SLPs who participated in an online rating task. Participants rated 11CSVs illustrating a standardized patient performing the 3-ounce WSC. Non-expert participants received training, feedback on their performance, on-demand definitions, and unlimited CSV reviews. Expert participants received training with no feedback on their performance and unlimited CSV reviews. On-demand definitions were not available. Non-expert participants completed eight demographic questions related to work setting, experience, and clinical practices. Results: Non-expert mean accuracy (M= 90.06, SD = 8.45) was significantly higher than expert mean accuracy by a mean difference of 9.59, 95% CI [8.23 to 10.95], $t(149) = 13.89$, $p < .001$. Non-expert and expert intra-rater reliability revealed 91% and 98% overall proportion of agreement, respectively. Non-expert interrater reliability revealed "good" agreement ($\kappa = .69$, 95% CI [.692, .703], $p < .001$). Expert interrater reliability revealed "good" agreement ($\kappa = .65$, 95% CI [.570, .727], $p = .000$). Non-expert demographics were not statistically significant predictors for CSV accuracy. Conclusions: Expertise did not influence rating accuracy or reliability. Training with knowledge of performance and on-demand definitions review improved CSV rating accuracy. Demographics did not predict rater performance as reported by previous investigations. Updated WSC interpretation guidelines, including training with feedback and expanded definitions, should be considered.

Keywords: dysphagia, aspiration, swallow screening, reliability, Yale Swallow Protocol, Water Swallow Testing

Dedication

Dedicated to my sister, Brittany. There is not a day that goes by…

Acknowledgments

This work was made possible by the support of so many people:

To my committee: Thank you for your support and guidance while allowing me to carve out my own research interests. A special thank you is extended to Dr. Kristine Galek, Dr. Thomas Watterson, and Dr. Tami Brancamp. Thank you for allowing me to navigate this journey in a new and meaningful way. I will always appreciate your kindness and candor. Thank you for this opportunity.

To my teachers, especially Mr. Jordan Phee, Mrs. Ruth Litterini, Dr. Shelley Von Berg, and Dr. Kerry Lewis: You have inspired me more than you'll ever know. Thank you for encouraging the ongoing pursuit of life-long learning.

To my family and friends-turned-family: I will never understand your unreasonable confidence in me, though I am immeasurably grateful for it. Thank you for knowing what I needed and when I needed it most: your grace, tough-love, compassion, and words of encouragement. I am forever indebted to you. Thank you will never cut it.

Table of Contents

List of Tables

List of Figures

Introduction
Speech-Language Pathologists Ratings of the Yale 3-Ounce Water Swallow Challenge: Accuracy, Reliability, and Clinician Demographics

Swallow function is a complex process requiring the precise coordination of more than 30 muscles and nerves (Matsuo & Palmer, 2008). An impairment in swallowing function, known as dysphagia, is estimated to occur in one of every 25 individuals in the United States (Bhattacharyya, 2014). This condition is known to occur in the setting of both primary medical conditions (cancer, respiratory disease, neurologic conditions) and in otherwise-healthy community-dwelling older adults (Clavé & Shaker, 2015).

Swallowing impairment is a critical issue with the potential for significant negative impact. Dysphagia is associated with negative health sequelae, including aspiration pneumonia, weight loss, increased length of hospital stay, reduced quality of life, and death (Garand et al., 2020). Despite the prevalence and significant consequences associated, dysphagia is underrecognized by patients and healthcare professionals (Clavé & Shaker, 2015). As such, individuals at-risk for dysphagia require rapid and accurate identification to minimize the anticipated negative dysphagia-related sequelae.

Speech-Language Pathologists (SLPs) are responsible for screening, evaluating, and managing oral-pharyngeal swallowing impairment. A review of the existing literature identified wide variation and lack of consistency among SLP dysphagia screening, evaluation, and treatment practices (Carnaby & Harenberg, Martino et al., 2009; Vose et al., 2018). Swallow screenings are designed to identify individuals at risk for dysphagia or aspiration using a pass-fail criterion. Water Swallow Testing (WST) is a validated swallow screening tool. The Yale Swallow Protocol, a widely accepted WST, includes a 3-ounce Water Swallow Challenge (WSC). The Yale 3-Ounce WSC directives require

the rapid and sequential drinking of water without interruption. The administration and interpretation guidelines of the Yale 3-Ounce WSC are clear and simple and are used by SLPs in various clinical settings. Little evidence exists regarding SLP adherence to protocol interpretation guidelines. Therefore, little is known about the accuracy and reliability of SLPs interpreting the Yale Swallow Protocol 3-Ounce WSC.

Literature Review

Screening for Swallowing Impairment

The American Speech-Language and Hearing Association (ASHA) defines swallow screening as "*...a pass or fail procedure to identify individuals who require a comprehensive assessment of swallowing function or a referral for other professional and medical services*" (ASHA, n.d.).

Swallow screening is intended to identify individuals requiring an SLP assessment (ASHA, n.d; Donovan et al., 2013; Perry, 2001; Suiter et al., 2020). Swallow screening differs from comprehensive assessments not capable of determining swallow pathophysiology or providing treatment-related recommendations (e.g., compensatory maneuver, modified diet texture, rehabilitative exercises).

Despite the focused purpose, a review of the literature identified several benefits to screening for swallowing impairment. Swallow screening has been identified as a concise and cost-effective strategy to identify individuals requiring specialized comprehensive swallow evaluation. The identification of dysphagic individuals has been associated with a reduction in the negative impact of dysphagia-related sequelae (Hinchey et al., 2005; Wangen et al., 2019).

A 2005 prospective study conducted by Hinchey et al. examined the development of hospital-acquired pneumonia in post-stroke patients. Their findings indicated that a formal dysphagia screening protocol reduced the rate of hospital-acquired pneumonia when compared to rates of hospitals without a formalized post-stroke swallow screening protocol.

Wangen and colleagues (2019) examined aspiration-related mortality of acutely ill hospitalized patients. Their findings identified reduced mortality rates following the initiation of a program to screen for aspiration risk. Swallow screening was also found to minimize unnecessary oral intake restrictions and was associated with a safe return to oral intake (Leder et al., 2012).

Leder et al. (2012) examined the clinical utility of recommending oral intake based on the outcome of a formalized swallow screening protocol. To accomplish this, the authors monitored the oral intake of hospitalized patients for 12-24 hours following the successful completion ("pass" result) of a 3-ounce water swallow challenge. Their results indicated that patients that passed a swallow screening could safely continue oral intake without the need for additional evaluation.

Swallow Screening Psychometrics: Sensitivity, Specificity, Validity, & Reliability

Screening instruments should be valid, reliable, and sensitive (Suiter, 2018). The diagnostic accuracy of a swallow screening instrument is dependent upon psychometrics to identify patients accurately and consistently with dysphagia/aspiration (i.e., sensitivity) from patients without dysphagia and/or aspiration (i.e., specificity). When related to a swallow screening, sensitivity refers to the closeness with which a positive result (e.g., coughing observed during a swallow screening) and a positive presence of the condition (e.g., dysphagia/aspiration confirmed by an instrumental evaluation) occurs. Specificity, as it relates to swallow screening, refers to the closeness with which a negative result during swallow screening (e.g., no coughing observed) corresponds with a negative result (e.g., normal swallow function confirmed by an instrumental evaluation).

Ideal screening instrument psychometrics would deliver both high sensitivity and specificity. A screening tool that accurately includes those with high dysphagia/aspiration likelihood and excludes individuals with a low dysphagia/aspiration likelihood is optimal. This is not possible, however, as a strength observed in one metric results in the inverse in the remaining metric (Simundic, 2009). Screening instruments with high sensitivity are preferred over high specificity, citing a clinical preference to obtain inaccurate positive screening results over inaccurate negative results (Suiter, 2018). The inaccurate identification of a truly non-dysphagic patient with a positive screening result is preferable to failing to identify truly dysphagic individuals.

Validity refers to the instrument's accuracy in measuring the intended target (Jacobsen, 2017; Shadish et al., 2002). For the purpose of this study, a screening purported to measure dysphagia/aspiration should, in fact, measure aspiration/dysphagia. A swallow screening instrument that reports the ability to identify aspiration but was found to measure dysphagia would not be valid. Conversely, a screening instrument that reports the ability to identify and was found to identify aspiration when compared to the reference test would demonstrate validity.

Reliability, as it relates to the current investigation, refers to the consistency of a rating or response. Reliability in this study measures the judgment/response agreement of multiple raters and the consistency measurements made by a single rater across multiple time points (Frowen, Cotton, & Perry, 2008). The reliability of SLP judgments has been examined. In 2000, McCullough et al. examined the inter- and intrajudge reliability of SLPs examining patient performance during clinical evaluation tasks. Their findings reported SLP reliability of judgments in less than 50% of the clinical evaluation tasks.

These findings were attributed to both the clinical examination task itself, variation among providers, and the limited stability of patient performance over time. While patient variability cannot be controlled for, McCullough et al., (2000) suggested that the variability of clinician judgments may be minimized by operationalizing the process and providing SLP training.

Factors Impacting SLP Reliability

Training Effects: The Intersection of Training, Practice, Feedback, and Anchors

The role of training and its relationship to reliability is well established in the literature (Clain et al., 2021; Hind et al., 2009; Martin-Harris et al., 2008; Scott, Perry,& Bench, 1998). A review of the literature identified that a combination of group discussions, training with consensus, feedback on performance, and clear definitions have emerged as efficacious training effects to optimize rater reliability.

Scott, Perry, and Bench (1998) noted the positive impact of rater training on the reliability ratings of SLPs completing modified barium swallow studies. The method for their investigation utilized a 5-point rating scale for SLP judgments made during video fluoroscopic swallowing studies. The investigation included three treatment conditions: individual use of the scale without experience, individual use of the scale with conferring with other participants, and individual use of the scale after experience. The authors reported that SLP ratings completed in a group format with training and consensus prior to ratings resulted in superior reliability when compared to the two other treatment conditions. Timing of observations, bolus consistency, image quality, and task complexity were also identified as additional factors impacting SLP reliability.

The positive impact of training was echoed by Hind et al. (2009), who examined the accuracy of SLPs when assigning Penetration-Aspiration Scale (Rosenbeck et al., 1996) ratings. Their group determined that SLPs that underwent training with competency assessment demonstrated more reliable judgments when compared to the judgments of clinicians that did not receive training.

Clain et al. (2021) examined the psychometric properties of the Modified Barium Swallow Impairment Profile and found that MBSImP-trained SLPs achieved good reliability across a variety of demographic characteristics. Their research suggested that high-quality rater training may be more influential than rater demographic characteristics or practices.

The opportunity to practice the desired stimulus as a reliability-enhancing strategy has also been supported in the literature. Lee, Whitehill, & Ciocca (2008) identified that the opportunity to practice relevant stimuli, with or without feedback on performance, yielded a positive effect on the reliability of rater judgments.

Given the positive impact of rater training on reliability measures, contemporary works by Martin-Harris et al. (2008) required accurate and reliable rater judgments while validating the Modified Barium Swallow Impairment Profile (MBSImP) scoring protocol. To achieve this, Martin-Harris et al. (2008) engaged SLPs in multiple individual and group training sessions targeting the application of the physiological components of the MBSImP. SLPs scored multiple MBSImP videos, and their ratings were compared against the standard (Martin-Harris). A consensus was reached when a passing reliability score (≥ 80% accuracy) was obtained; the authors reported high reliability following the standardized training.

The known benefit of training can be observed clinically, as multiple therapeutic programs (McNeill Dysphagia Therapy Program, Lee Silverman Voice Therapy, SpeakOUT!) require standardized training with an assessment of clinician knowledge as a strategy to optimize reliability.

Anchors have also been identified as a strategy to improve reliability. This strategy was described in Goldstone's (1998) work discussing the mechanisms of perceptual learning. Roughly translated, this strategy described raters' use of their own repeated experiences to formulate an internal standard for perceptual learning. Individual internal standards understandably vary from rater to rater and may not remain stable over time, resulting in wide variety and poor reliability of rater judgments.

The application of the concept of anchors from perceptual voice studies to other perceptual tasks could conceivably be accomplished by using external standards/anchors to serve as references for comparison. Existing literature regarding anchors has focused on perceptual voice and resonance judgments. Although anchors have not been discussed in the swallow literature, it is reasonable to consider the benefit of anchors as definitions for perceptual judgments as a strategy to promote objectivity while evaluating subjective tasks. Prior research has confirmed the positive impact of clear definitions on rater reliability (Stoeckli et al., 2003).

Anchors and training effects were examined as mechanisms of perceptual learning by Chan and Yiu (2002). Here, the authors applied a stimulus-response-feedback-stimulus training program to examine training effects (feedback and no-feedback) and voice stimuli (synthesized and natural) on the reliability of rater judgments. All participant groups in this study received training. Chan and Yiu reported that training and

auditory anchors, regardless of feedback, led to higher accuracy and reliability. While it was not examined in their study, the authors questioned if the application of anchors alone without training would impact rater reliability.

Task Complexity

The evaluation of swallow function is inherently complicated, as SLPs are required to make multiple judgments during the evaluative process (Scott, Perry, & Bench, 1998). Scott, Perry, & Bench (1998) found that task complexity may impact interrater reliability and recommend the simplification of procedures to optimize psychometrics. In their 1988 study, Ekberg et al. examined interobserver variability of clinician judgments during videofluoroscopy. When examining agreement among practicing radiologists, their research indicated that certain features observed during fluoroscopy yielded higher levels of agreement, while other stimuli results in poorer agreement among raters. Specifically, features that were easier to observe cineradiographical were related to stronger concordance of judgments. “Easy” features to identify included absent pharyngeal constriction, airway invasion, and normal swallow function.

Research also indicates that the complexity of a clinical task may impact rater performance (Martin-Harris, 2015; Scott Perry & Bench, 1998; Vose et al., 2018;). Martin-Harris (2015) previously discussed that the functional complexity of individual patients may be aided by the use of standardized evaluation protocols. The concept of task complexity was also present in Vose and colleagues’ 2018 survey of SLP clinical decision-making. This investigation presented stimuli of easy, moderate, and complex levels of difficulty that were evaluated by SLP participants. Their results found that

complexity and agreement were inversely related; agreement between SLP recommendations decreased as video complexity increased.

Rater Demographics: Experience & Clinical Practices

It has been reported rater demographic characteristics may impact reliability measures, however, the research regarding the significance of this is mixed.

Zarkada and Regan's 2018 investigation of interrater reliability on swallow evaluation judgments determined that raters' clinical experiences influenced performance. In this study, raters with more experience demonstrated significantly higher interrater reliability when compared to the performance of less experienced raters.

The influence of experience was also identified by Lewis, Watterson & Houghton (2003), who found that raters with more experience demonstrated higher levels of agreement than the less experienced raters when asked to complete a listening rating task.

Rater experience has been shown to impact clinical judgments (Scott, Perry, & Bench, 1998, p. 227) however the particular impact on reliability was not reported in this study. Ekberg et al. (1988) identified that interrater agreement was correlated experience with their findings that participants with more experience demonstrated fewer disagreements on fluoroscopy observations.

In 2021, Clain et al. examined the psychometric properties of the Modified Barium Swallow Impairment Profile, a standardized approach to the analysis of the Modified Barium Swallow Study. Their findings reported that more experienced raters demonstrated stronger reliability when compared to raters with fewer years of experience,

though overlapping confidence intervals for all raters warranted exclusive future investigation of the experience-reliability relationship.

SLP years of experience are not the only demographic characteristic for consideration. Vose et al. (2018) determined that clinician practices influenced rater accuracy. In this study, clinicians that reported the use of -frame-by-frame review for modified barium swallow study interpretation demonstrated better performance than SLPs that did not report use of frame-by-frame analysis.

Swallow Screening Modalities

Non-Validated Screening Methods

Screening for swallowing impairment may be completed by means of non-validated and validated screening modalities. Non-validated swallow screening methods incorporate indirect and direct swallow screening approaches (see table 1). Indirect swallow screening refers to the connotation that the act of swallowing is not directly observed. Rather, this swallow screening method relies on non-swallowing tasks as a surrogate indicator for signs of swallowing impairment. Examples of indirect swallow screening methods include a review of the medical record (Mari et al., 1997) or the completion of a patient interview. Indirect swallow screening may implement a therapist-generated procedure or a facility-generated screening procedure (Etges et al., 2014).

Direct **s**wallow screening methods include the observation of patient behaviors during a given task. Direct swallow screening procedures may include oral bolus trials while monitoring for overt signs and symptoms of swallowing impairment, including coughing or throat clearing (Kidd et al., 1993; Daniels et al., 2000; Logemann et al.,

1999)., change to voice quality (Groves-Wright et al., 2010; Ryu et al., 2004; Warms & Richards, 2000), oxygen desaturation (Britton et al., 2018; De Groof et al., 2004; Exley 2000; Leder 2000; Marian et al., 2017; Wang et al., 2005; Zaidi et al., 1995), and cervical auscultation (Al Hawat et al., 2014; Borr et al., 2007; Dudik et al., 2018, Sarraf Shirazi & Moussavi, 2012; Stroud et al., 2002). Alternatives to offering oral bolus trials include oral motor skills (Daniels et al., 2000; Logemann et al., 1999; McCullough et al., 2001), examination of tongue strength (Lee & Choi, 2020) and gag reflex testing (Kidd et al., 1993).

Table 1

Non-Validated Swallow Screening Methods

Screening item	Author(s)
Medical history review for etiological risk categories	Mari et al. (1997)
Use of decision-making algorithms	Runions et al. (2004).
Evaluation of gag reflex or pharyngeal sensation	Kidd et al. (1993)
Overt signs of cough or other difficulty during oral bolus trial swallows with water	Daniels et al. (2000); Hughes & Wiles (1996); Kidd et al. (1993); Logemann et al. (1999); Nathadwarawala, Nicklin, & Wiles (1992);
Jaw-opening force test (JOFT)	Hara et al. (2014)
Phonation	Festic et al. (2016)
Cough	Addington et al. (1999); Curtis & Troche (2020); Guillén-Solà et al. (2013); Lee et al. (2014); Sato et al. (2012); Wakasugi et al. (2014)
Maximum Phonation Duration	Lim et al. (2020)
Tongue Strength	Lee & Choi (2020)
Pulse oximetry	Britton et al. (2018); De Groof et al. (2004); Exley (2000); Leder (2000); Marian et al. (2017); Wang et al. (2005); Zaidi et al. (1995)
Cervical Auscultation	Al Hawat et al. (2014); Borr et al. (2007); Dudik at al. (2018); Sarraf Shirazi & Moussavi (2012); Stroud et al. (2002)

Wet vocal quality (WVQ)	Groves-Wright et al., (2010); Ryu et al. (2004); Warms & Richards (2000)
Pitch elevation	Malandraki et al. (2011); Rajappa et al. (2017)
Clinician surveillance of oral motor praxis, voice pitch change, swallow bolus trials	Daniels et al. (2000); Logemann et al. (1999); McCullough et al. (2001)

Research has examined the efficacy of non-validated swallow screening methods. Repeatedly, the diagnostic accuracy of non-validated swallow screening methods were found to poorly identify the presence of swallowing impairment (McCullough et al., 2001). Research has also found that the inclusion of validated screening measures to detect aspiration has significantly improved the ability to accurately and reliably identify individuals likely to demonstrate aspiration (Brodsky et al., 2016). SLPs have been encouraged to implement validated screening methods over independently established "facility-specific" protocols, considering the diagnostic weaknesses of the latter (Donovan et al., 2013).

Validated Screening Methods

Patient-Reported Outcome Measures. Validated questionnaires, surveys, and scales allow patients to quantifiably measure their dysphagia-related quality of life. A brief description of dysphagia-related quality of life PROMS validated for heterogeneous populations are presented below (see table 2).

The use of PROMs as a swallow screening method may appear counterintuitive considering the absence of a swallow task. However, the EAT-10 (Belafsky et al., 2008; Cheney et al., 2015) and the SWAL-QOL/ SWAL-CARE (McHorney et al., 2002) underwent validation to predict swallow dysfunction. While an EAT-10 (Belafsky et al., 2008; Cheney et al., 2015) score of 3 or greater is considered “abnormal” for swallow-

related quality of life, scores of 15 or greater predicted aspiration with a demonstrated sensitivity of 71%. Likewise, SWAL-QOL (McHorney et al., 2002) scores specific to quality of life have also been correlated to bolus flow characteristics on instrumental swallow assessment.

While the use of validated PROMs as a swallow screening method may be preferable as an alternative to the use of non-validated screening methods, PROMs reliance upon patient-reported symptoms presents several known challenges. Suiter (2020) expressed that PROMs assume intact cognitive-linguistic function and the effect of altered cognitive-linguistic function for patient-reported dysphagia symptoms is unknown. Prior research has also found that patients often underreported the severity of their swallowing symptoms (Ding and Logemann, 2008). The extent to which the aforementioned challenges impact PROMs psychometric integrity is not known.

Table 2

Patient-Reported Outcome Measures (PROMs)

Instrument	Features
Swallow Quality of Life Questionnaire (SWAL-QOL, & SWAL-CARE) (McHorney et al., 2002)	44-item SWAL-QOL addressing ten quality of life areas 15-item SWAL-CARE addressing quality of care and satisfaction Five-point scale with lower scores indicating higher level of impairment SWAL-QOL scores have been correlated to bolus flow characteristics on instrumental swallow assessment.
The Sydney Swallow Questionnaire (SSQ; Wallace et al., 2000)	17-item survey addressing symptom severity in oral-pharyngeal dysphagia. Total score out of 1,700.
Eating Assessment Tool (EAT-10; (Belafsky et al., 2008; Cheney et al., 2015)	Ten-item patient-administered survey established to quantify dysphagia severity. A five-point scale (zero-four) in which higher scores indicate increased symptom severity. Scores greater than three are considered "abnormal." 2-minute administration time.
Dysphagia Handicap Index (DHI; Silbergleit et al., 2012)	25-item questionnaire measuring handicapping effect of dysphagia.

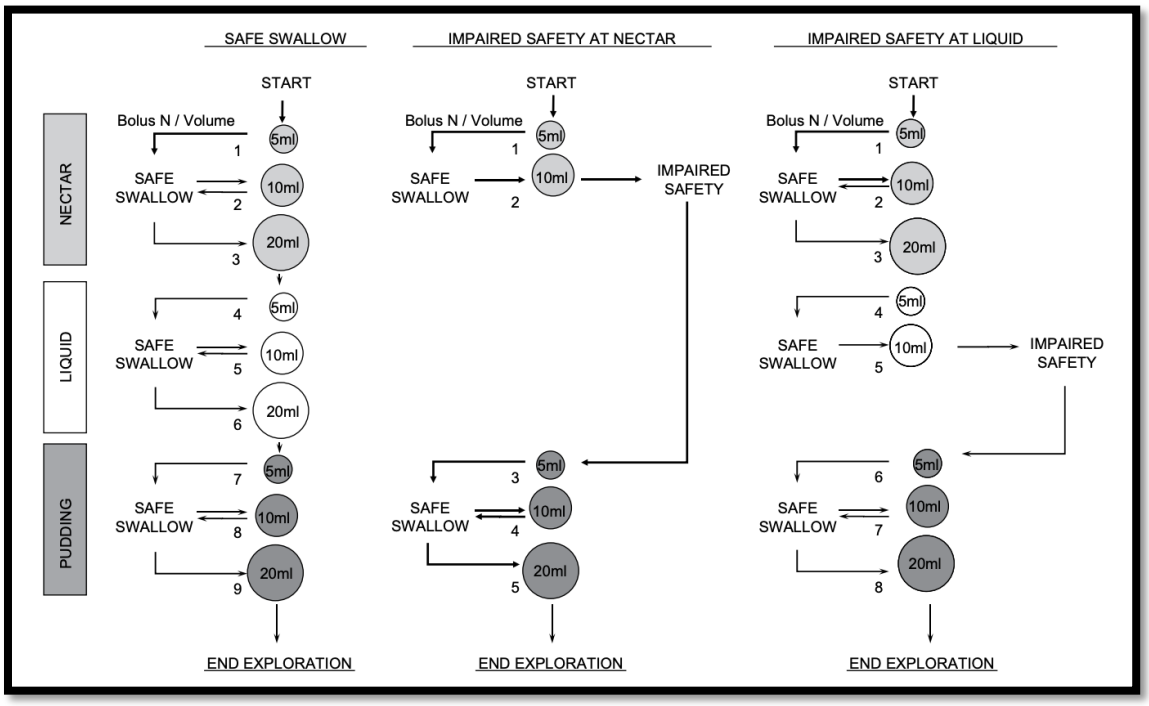
Behavioral swallowing screening instruments. Behavioral swallow screening instruments refer to administration and interpretation protocols validated against a reference test (fluoroscopy, endoscopy). A review of the literature identified two behavioral screening instruments validated for use with a heterogenous patient population: Volume-Viscosity Swallowing Test (V-VST; Clave et al., 2008; Rofes et al., 2012) and Yale Swallow Protocol (YSP; Suiter & Leder, 2008; Suiter & Leder, 2014; Ward et al., 2020).

The Volume-Viscosity Swallow Test (V-VST; Clave et al., 2008; Rofes et al., 2012). is a screening instrument validated to identify dysphagia and aspiration in neurologic (*N*=85) patients and community-dwelling (*N*=254) individuals. This screening protocol includes the administration of three bolus amounts and viscosities with clinician monitoring for indication of impaired swallow safety of efficacy (Figure 1). Impaired swallow safety is defined as the presence of a cough response, oxygen desaturation measured by pulse oximetry, and/or a change in vocal quality. Impaired swallow efficacy was defined as piecemeal deglutition and oropharyngeal residue. Videofluoroscopy was performed as the reference test.

The V-VST identified impaired safety (oxygen desaturation, cough, vocal quality change) with 88.2% sensitivity which increased to 100% sensitivity for aspiration. The V-VST identified impaired swallow efficacy (piecemeal deglutition, oropharyngeal residue) with 88.4% sensitivity.

Figure 1

Volume-Viscosity Swallowing Test Administration Algorithm (Clave et al., 2008)



Research detailing the SLPs ability to identify swallowing impairment accurately and reliably from the bedside is encouraging; however, the authors' findings present several challenges that should be considered. Most notable is the author's identification of laryngeal vestibule penetration as "unsafe" swallow behavior. Many normal and healthy individuals experience episodes of penetration during oral intake (Rosenbeck et al., 1996). The strong sensitivity of this screening instrument may be inflated by broad screening fail criteria. A description of penetration depth and the patient's response to the penetration episode may present more clinically meaningful information.

Regarding the screening stimuli, Clave et al. (2008) selected small bolus volumes and "safer" viscosities in an attempt to minimize risk or harm to patients in the event that swallowing impairment was likely. Unfortunately, subsequent literature has identified

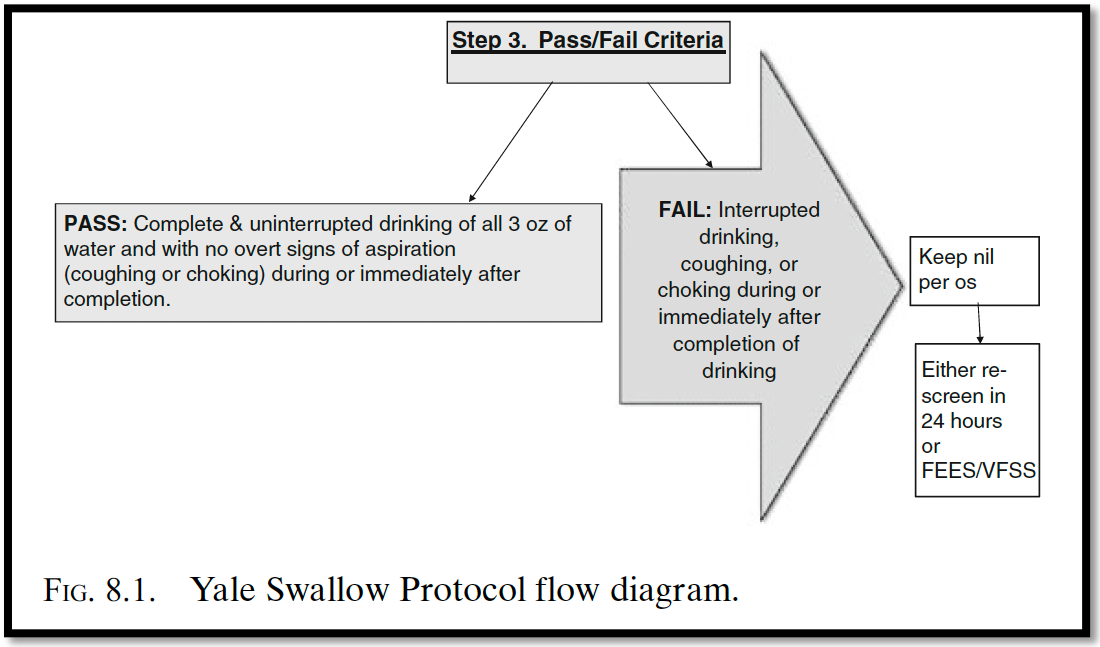
that small bolus volumes were associated with lower sensitivity to aspiration when compared to larger bolus volumes (Brodsky et al., 2016; Leder et al., 2011).

Methodologically, the authors provided a limited report of training, consensus, and reliability measures. This information is paramount to a "good" screening tool to ensure protocol adherence and maintain instrument validity. The absence of this information could reasonably degrade the psychometric validation of the V-VST when implemented by SLPs in clinical practice.

The Yale Swallow Protocol (YSP; Suiter & Leder, 2008, Suiter & Leder 2014, Ward et al., 2020). is a screening instrument validated to identify aspiration accurately and reliably in a heterogenous acute (N = 3000) and post-acute (N=240) patient populations. The YSP has three components: a cognitive screening, an oral mechanism examination, and a 3-ounce water swallow challenge. The water swallow challenge is the only scored component of the YSP. During the 3-ounce water swallow challenge (WSC), the patient is instructed to drink three ounces of water in succession without interruption. Patient performance is scored on a pass/fail criterion. A pass result is awarded when the patient drinks all three ounces of water in succession without interruption or coughing during or immediately following the 3-ounce water swallow challenge. A fail result is assigned if the 3-ounce WCS is interrupted or if a cough response occurs either during or immediately following the 3-ounce water swallow challenge (see figure 2).

Figure 2

Yale Swallow Protocol 3-Ounce Water Swallow Challenge Interpretation Guidelines (Leder & Suiter, 2014)



Verified by endoscopy as the reference test, a recent revalidation of the YSP reported 95.4% sensitivity and 66.9 % specificity in the post-acute setting (Ward et al., 2020).

The YSP offers many advantages for clinicians. First, the YSP reports higher sensitivity when compared to other screening instruments. The YSP reports lower specificity when compared to other screening instruments. This is not considered a disadvantage, as a higher sensitivity and lower specificity are preferable to a low sensitivity and a high specificity (Suiter et al., 2020). The robust sample size ($N = 3000$) and validation for clinical implementation with a diverse patient population offers an additional benefit to clinicians.

Patient performance on the YSP is determined by scoring of a single component, the 3-ounce water swallow challenge (WSC). Research indicates that the 3-ounce volume of water provided higher sensitivity to aspiration compared to other water swallow tests that implement smaller volumes of water (Brodsky et al., 2016).

The YSP also presents several issues for consideration. The YSP reported 100% interrater agreement with the blinding of raters during validation studies. Specific training and reliability-based information were not reported. As such, it is reasonable to question if all SLPs can preserve the YSP's high level of psychometric validation in the absence of training instructions or a competency evaluation.

A central concern regarding the YSP interpretation guidelines related to the absence of operational definitions for significant terms. The YSP 3-ounce WSC interpretation guidelines instruct that a "fail" response is assigned following the observation of interrupted drinking or coughing that is observed during or immediately following the conclusion of drinking. Unfortunately, the YSP does not define the length of time which may be considered "immediate." Additionally, the authors appear to both fail to operationally define "coughing or choking" and use "choking" as a surrogate term for a laryngeal response. It is possible that an absence of operational definitions may contribute lower interrater reliability and/or impact SLPs accuracy when assigning a pass/fail response.

Clinical Implementation of the Yale Swallow Protocol (YSP)

The YSP has emerged as a reliable aspiration screening tool with validation for wide clinical application. Research detailing the clinical application of the YSP outside of the original validation is limited.

Suiter, Sloggy, & Leder's (2014) prospective double-blind YSP validation study reported 100% reliability of between the two evaluating SLPs with reported familiarity in YSP administration. A description of prior training or consensus information was not reported. Warner et al. (2014) detailed the accuracy of the registered nurses (RNs) performing the YSP. In this study, the authors investigated the reliability and accuracy of hospital-based RNs' administration and interpretation of the YSP. 52 RNs performed the YSP across 101 hospital patients. The judgments of two SLPs were used at the standard; SLPs demonstrated 100% intra- and interrater agreement. Findings indicated that, following a web-based training and competency demonstration, RNs demonstrated 98% accuracy for correct interpretation of the Yale Swallow Protocol when compared to standard. Details and method of training for the two SLPs were not reported.

Ward et al. (2020) validated the YSP for use with the post-acute care patient population. In this study, the YSP was administered and interpreted by 66 skilled nursing facility (SNF) SLPs and compared against endoscopy that served as the reference test. SNF SLPs did not receive special training and were provided only the standard written YSP administration and interpretation protocol. The authors reported endoscopist blinding to the SNF SLP's ratings. Reliability of the SNF SLP and endoscopist SLP judgments was not reported.

Nielsen, Gow, and Svenningsen's 2021 study translated and adapted the YSP from English into the Danish language. This study did not examine YSP administration or interpretation and cannot be compared for accuracy or reliability. It is noteworthy to mention, however, that the authors identified training as necessary for nurses to reliably screen patients for swallow impairment.

Finally, Garand et al. (2021) examined YSP validity and accuracy when identifying aspiration risk in patients with motor neuron disease (MND). The sensitivity and specificity of the YSP was highlighted. SLP training and reliability measures were not reported.

It should be mentioned that swallow screening may be completed by allied health care professionals (i.e., MD, RN, PA-C). This practice is common when caring for patient populations with known risks for swallow impairment (i.e., stroke, extubation; Perry, 2001). The benefits of this practice may include the timely identification of patients with suspected swallow impairments, the ability to initiate or withhold oral medications and nutrition, and the avoidance of extended *nil per os* (NPO) while awaiting SLP consultation (Edmiaston et al., 2014; Suiter et al., 2020). Research regarding RNs YSP proficiency has demonstrated a high degree of accuracy and reliability with proper training was provided (Anderson et al., 2016; Campbell et al., 2016; Weinhardt et al., 2008).

Existing studies have focused on the ability of the YSP to identify aspiration with little mention of SLP training. Leder & Suiter (2014) instructed that the YSP should be administered by "trained healthcare professionals" without further clarification or instruction on training. Despite the reported strengths of training, formal YSP training or competency assessment is not available nor required for SLPs.

SLP Swallow Screening Practice Patterns

The goals and benefits of swallow screening are well established by the literature (ASHA, n.d.; Brodsky et al., 2016; Suiter, 2018). A review of the existing research examining SLP dysphagia clinical practices identified the lack of standardization as a

central issue (Carnaby & Harenberg, 2013; McCullough et al., 2000; Vose et al., 2018). Literature describing the practice patterns of dysphagia clinicians has identified significant variation regarding dysphagia evaluation (Martino et al., 2009), clinical decision making (Vose et al., 2018), and rehabilitation practices (Carnaby & Harenberg, 2013). This variation was so substantial, in fact, that Carnaby & Harenberg (2013) reported a true lack of consensus in dysphagia care.

This variation in clinical practices is likely multifactorial. One probable contributor to this variation might be related to SLPs use of non-standardized and non-validated methods. Carnaby & Harenberg's (2013) examination of usual care in dysphagia rehabilitation found that SLPs reported frequent implementation of self-generated evaluation methods rather than published evidence-based practice. Roberts et al. (2020) identified multiple barriers to implementation of evidence-based practices; barriers were attributed to limited time, a lack of support from superiors, and lack of available evidence among providers.

The use of non-validated and non-standardized methods are of low clinical yield given their poor diagnostic accuracy. Prior research by McCullough et al. (2001) found that clinical bedside swallow evaluation methods poorly identified the presence of swallowing impairment. Specifically, SLP reliance upon signs/symptoms observed during the non-instrumental clinical swallowing evaluation as clinical indicators for dysphagia or aspiration truly identified swallowing impairment in 50% of opportunities. As such, SLPs should implement validated screening methods over independently established "facility-specific" protocols, considering the diagnostic weaknesses of the latter (Donovan et al., 2013). The inclusion of validated screening measures to detect

aspiration has significantly improved SLPs ability to accurately and reliably identify individuals likely to demonstrate aspiration (Brodsky et al., 2016).

For SLPs that implement validated swallow screening methods, the issue of SLP training bears discussion. At present, validated swallow screening instruments do not provide nor require SLP protocol training prior to administration. An absence of protocol training is disadvantageous for several reasons. A lack of protocol training and competency assessment relies upon inherent SLP training and skills to administer accurately and reliably said screening protocol. Without training and competency assessment, clinicians may unintentionally deviate from the protocol guidelines, thereby degrading its psychometric validation. Clinicians utilizing a validated swallow screening instrument should consider the appropriate clinical population and the respective procedures and endpoints. Clinicians must ensure that the screening instrument is indeed valid for use with the respective clinical population to ensure valid and reliable results.

Research Problem

Dysphagia is a critical issue with the potential for negative health consequences, including pneumonia, weight loss, malnutrition, reduced quality of life, increased healthcare spending, and death (Garand et al., 2020). To minimize the likelihood and/or severity of these negative health consequences, SLPs are responsible for screening, evaluating, and managing oral-pharyngeal swallowing impairment.

A robust body of literature supports the implementation of validated swallow screening instruments to accurately identify individuals at risk for swallowing impairment. Validated screening methods have demonstrated improved bedside identification of individuals requiring complete SLP swallowing evaluation. Previous

research has identified wide variation and lack of consistency among SLP dysphagia screening, evaluation, and treatment practices (Carnaby & Harenberg, 2013; Martino et al., 2009; Vose et al., 2018). This variation has resulted in poor reliability of SLP judgments.

The YSP has emerged as a reliable swallow screening instrument validated for use across a large, heterogeneous patient population. Prior research has validated the YSPs ability to accurately and reliability identify swallowing impairment characterized by aspiration of thin liquids. To date, little is known about the clinical implementation and interpretation of the Yale Swallow Protocol as dysphagia screening among practicing SLPs.

After participant training, this study examined the accuracy, reliability, and demographic characteristics of SLP judgments across a variety of clinical scenario videos (CSVs). This study was the first to extend the use of clinical scenario videos to water swallow testing.

Study Aim

The investigation aimed to increase our understanding of SLP swallow screening practices. The investigator accomplished this in two ways. First, by examining the accuracy and reliability of participant responses across different novel clinical scenario videos. Second, the investigator explored the relationship between SLP clinical scenario video rating accuracy scores and their respective demographic features.

Method

Phase I: Pilot Investigation

A pilot study was conducted to determine feasibility, formalize study questions, optimize clinical scenario videos (CSVs), and establish expert rater performance for comparison in the final investigation. The pilot study was reviewed by the University IRB (Influence of Rater Experience on Water Swallow Testing Interpretation Protocol 1838584-1, exempt).

The investigator proposed four research questions to be addressed during the pilot study.

Research Questions

Research Question 1. To what extent can experts accurately judge 3-ounce water swallow challenge videos?

Null hypothesis: There is no difference in participant mean accuracy and the test value (75).

H_0: $\mu = 75$.

Alternate hypothesis: Mean accuracy and the test value are not equal. H_A: $\mu \neq 75$.

Research Question 2. To what extent does rater accuracy of 3-ounce water swallow challenge ratings differ among delivery method (cup, straw)?

Null hypothesis: There is no difference in mean accuracy of water swallow interpretations among delivery method (cup, straw). H_0: $\mu cup = \mu straw$

Alternate hypothesis: There is a difference in mean accuracy of water swallow interpretations among delivery method (cup, straw). H_A: $\mu cup \neq \mu straw$

Research Question 3. To what extent can experts reliably judge 3-ounce water swallow challenge videos?

Null hypothesis: Agreement between raters is no different than chance agreement. H_0: $\kappa = 0$.

Alternative hypothesis: Kappa (κ) coefficient is different from zero/chance agreement. H_A: $\kappa \neq 0$

Research Question 4. To what extent does rater reliability of 3-ounce water swallow challenge ratings differ among delivery method (cup, straw)?

Null hypothesis: Agreement between raters among delivery method (cup, straw) is no different than chance agreement. H_0: κ cup= κ straw.

Alternative hypothesis: Kappa (κ) coefficient for delivery method (cup, straw) is different from zero/chance agreement. H_A: κ cup $\neq$ κ straw.

Procedure

The research project was a non-experimental cross-sectional correlational design study. The investigator did not actively manipulate variables, as participant groups are based on self-reported demographic information.

Participants and Recruitment/Sampling Procedure

Ten participants were identified by the investigator as expert raters. Each expert rater was selected secondary to their recent publications relevant to this project. Experts included nine SLPs and one Registered Nurse (RN), Ph.D. The nine SLPs were employed in a variety of work settings including academic medicine/higher education, voice and swallowing clinics, and swallowing evaluation providers. The RN, Ph.D. participant is employed as a critical care nurse in Denmark.

Following Institution Review Board approval (Influence of Rater Experience on Water Swallow Testing Interpretation, Protocol number 1838584-1), the ten experts received an electronic mail invitation to participate in an anonymous 35-item online rating task examining clinical decision making.

Instrument

A University of Nevada, Reno School of Medicine standardized patient was hired to portray a patient completing a 3-ounce WSC. The investigator generated 17 novel CSVs to represent possible patient behaviors observable during a 3-ounce water swallow challenge (WSC, presented in Table 3). The clinical scenarios represented by the CSVs were created from the investigator's own clinical experiences and observations during YSP administration. CSVs were created by the investigator to allow for the examination of task complexity and YSP interpretation. Of note, the YSP does not delineate protocol interpretation beyond pass-fail criteria.

CSVs illustrated a clinical scenario demonstrated across the cup and straw delivery methods. To clarify, CSVs were generated in a paired format, meaning that a single clinical scenario was performed for both the cup and straw delivery method to allow examination of delivery method as a research variable.

The investigator trained the standardized patient to perform a 3-ounce WSC for each clinical scenario and delivery method.

Table 3 *Clinical Scenario Videos Presented for Expert Rater Interpretation*

Delivery method				
Cup		Straw		
Video number		Video number		
				Working standard response
1.	Uninterrupted drinking	10.	Uninterrupted drinking	Pass
2.	Delayed cough			Fail
3.	Uninterrupted drinking			Pass
4.	Repeat instruction	11.	Repeat instruction	Pass
5.	Throat clear	12.	Throat clear	Pass
6.	Interrupted drinking & coughing	13.	Interrupted drinking, & coughing	Fail
7.	Interrupted drinking & throat clearing	14.	Interrupted drinking & throat clearing	Fail
8.	Interrupted drinking, no cough nor throat clear	15.	Interrupted drinking, no cough nor throat clear	Fail
9.	Uninterrupted drinking with immediate cough	16.	Uninterrupted with immediate cough	Fail
		17.	Uninterrupted with immediate cough	Fail

Standardized patient performance was video recorded and edited by the investigator. Raw videos were trimmed to eliminate excess video before and after the 3-ounce water swallow challenge. Additionally, separate video image angles (anterior-posterior, lateral) were combined into a single side-by-side video image. Finalized videos were uploaded to the investigator's unlisted YouTube channel (Morrissey, n.d.), converted into HTML format, and embedded into Qualtrics Experience Management Software (Qualtrics, Provo, UT) survey template.

Instrument Modules. Based upon evidence for standardized clinical training and consensus, the instrument included three modules: learning, training, and reliability. The three-module format was modeled upon the Modified Barium Swallow Impairment Profile (MBSImP; Martin-Harris et al., 2008), a standardized protocol for interpreting video fluoroscopic swallow studies. Martin-Harris reports that the MBSImP three-module

design was self-generated using a “logical approach logical approach to other competency-based learning in our clinics” (personal communication, June 22, 2022). Although Martin-Harris did not follow adhere to a particular theoretical model during MBSImP generation, the integration of psychometrically sound training practices (training practice with feedback, and competency assessment) are readily apparent.

The Learning Module. presented a 2-minute video tutorial providing participant training on the administration and interpretation of the 3-ounce water swallow challenge task of the YSP. The video tutorial included verbal instruction and visual supports detailing the pass/fail criteria and YSP definitions. Participants received the following instruction:

> “The Yale Swallow Protocol is a validated screening method designed to reliably and accurately identify aspiration risk. In this video, you will learn to administer and interpret a single component of the Yale Swallow Protocol; the 3-ounce WSC. It’s important to note that the Yale Swallow Protocol is one of several water swallow tests. Each test presents its own administration and interpretation guidelines. For the purposes of this study, we are looking only at the Yale Swallow Protocol. The Yale Swallow Protocol has three components: a cognitive screening, oral mechanism exam, and the 3-ounce WSC. The patient is provided a cup containing three ounces of water and provided the following instructions “drink this water, slow and steady, without stopping.” Performance is interpreted by observing behavior during and immediately following the conclusion of drinking. The Yale Swallow Protocol is a screening measure and interpretation is stratified into pass/fail results. According to the flow diagram presented in the

> Yale Swallow Protocol, pass criteria includes complete and uninterrupted drinking of all 3 oz of water and with no overt signs of aspiration during or immediately after completion. Fail criteria is described as interrupted drinking, coughing, or choking during or immediately after completion of drinking. It is important to highlight that in Leder and Suiter's (2014) description, overt signs of aspiration are defined only by coughing or choking. As mentioned earlier, there are several other water swallow tests each with their own interpretation and administration guidelines. For the purpose of this study, we are considering only the Yale Swallow Protocol. Let's practice."

After reviewing the 2-minute instructional video, participants were presented with two practice CSVs. Participants reviewed each practice CSV and were prompted to apply the information presented in the instructional video tutorial to interpret the Yale Swallow Protocol's 3-ounce WSC. Participants evaluated standardized patient performance by applying one of three possible responses: *"The patient passed the water swallow challenge," "The patient failed the water swallow challenge,"* and *"I'm not sure."* Participants were provided with an "*I'm not sure*" response to avoid forcing of a response when the response was not known (Vose et al., 2018). Participants were permitted unlimited review of the instructional video (Stoeckli et al., 2003); however once answered, participants were not allowed to return to a previously answered CSV. Participants then transitioned to the training module.

The Training Module. required that participants apply the knowledge obtained during the instructional video across the 17 CSVs presented for the purposes of obtaining

accuracy and reliability data. Participants were presented with each CSV and were prompted to apply the information presented in the instruction video tutorial to interpret the YSP 3-ounce WCS.

Participants evaluated standardized patient performance by applying one of three possible responses: *"The patient passed the water swallow challenge," "The patient failed the water swallow challenge,"* and *"I'm not sure."* Participants were provided with an "*I'm not sure*" response to avoid forcing of a response when the response was not known (Vose et al., 2018). Definitions were available on demand. Participants were permitted unlimited review of the CSV (Stoeckli et al., 2003). Participants did not receive feedback on their performance. Once answered, participants were not allowed to return to a previously answered CSV. Participants then transitioned to the reliability module.

The Reliability Module. required that participants again apply the knowledge obtained during the instructional video. The reliability module presented participants with the same 17 CSVs presented in the prior training module. CSVs were presented in randomized order and participants were not made aware of the repetition.

Participants evaluated standardized patient performance by applying one of three possible responses: *"The patient passed the water swallow challenge," "The patient failed the water swallow challenge,"* and *"I'm not sure."* Participants were provided with an "*I'm not sure*" response to avoid forcing of a response when the response was not known (Vose et al., 2018). Definitions were available on demand. Participants were permitted unlimited review of the CSV (Stoeckli et al., 2003). Participants did not receive feedback on their performance. Once answered, participants were not allowed to return to

a previously answered CSV. Participants were then directed to the final instrument question.

The final question of the pilot rating task invited participants to provide feedback regarding clarity of task, quality of stimuli, barriers to completion, ease of completion, and/or issues related to functionality/software.

Operational Definitions and Scoring

The scoring of participants' CSVs judgments integrated the use of a reference standard and a working standard. According to Cook (2012), "a reference standard refers to the best available method for establishing the presence or absence of a condition of interest" (p. 111). A reference standard functions as the standard by which judgments are evaluated. Leder and Suiter's (2014) Yale Swallow Protocol interpretation definitions were selected by the investigator as the reference or gold standard following the proven sensitivity and specificity to accurately identify aspiration.

The investigator applied Leder and Suiter's (2014) reference standard to the novel CSVs presented in the pilot and final studies. This generated a "working standard" used for comparing participant CSV ratings and scoring for accuracy. A working standard is a measurement standard used for analysis and standardized against the reference standard (Ph. Eur., 2015). For the purpose of the study, the working standard is calibrated against a reference standard and is used to measure participant adherence to the reference standard (YSP). Participant responses were qualified by a working standard (AM) and scored for accuracy (Martin-Harris et al., 2008).

A working standard was applied for both expert and non-expert SLPs with consistent application of definitions and scoring during the pilot and final investigations and scored for accuracy.

Accuracy in this study is defined as participant ratings identical to those of the working standard as an established criterion. Participants were credited with correctly responding when their ratings matched the working standard. Accurate ratings received a "1" whereas inaccurate ratings received a "0." The participant score was computed by calculating the number of accurate responses divided by the total number of CSVs.

In this study, reliability over time and between raters is defined as the consistency of CSV judgments. The scoring procedures for both accuracy and reliability were completed in the same manner for each group.

When examining intrarater reliability, it is reported in the literature that researchers may re-attempt a small portion of initial judgments following time delay (Suiter et al., 2014). The anonymous nature of the study prevented response tracking and respondent identification, prohibiting future reassessment. All responses were obtained in a single encounter immediately following initial ratings. Intrarater reliability was assessed by an at-random repeat presentation of all previously viewed CSVs. Participants were not made aware of the repetition.

Data collection*:* 9 metadata variables including start date, end date, response type, progress, duration, completion, recorded date, distribution, user language (excluded from data extraction), 2 practice videos (excluded from data extraction), participant score, 34 videos, and one feedback question.

Data extraction: 34 CSVs (17 novel plus 17 repeated)

- 17 CSVs comparing the accuracy of participant ratings to the working standard
- 17 repeated CSVs comparing the consistency of one rater's responses across two separate rating opportunities

Data analysis: Participant responses were maintained by Qualtrics Experience Management Software (Qualtrics, Provo, UT). All responses are anonymous and no identifying information from the rater is associated with rating task responses. Rater responses cannot be traced to the participant. Anonymous Qualtrics rating task data was extracted, coded by the investigator, and imported into SPSS Statistics for Windows (Version 28.0.1) for analysis.

Statistical Analysis

A one-sample *t*-test will compare the mean accuracy of expert SLP ratings to experts. No precedent for percent exact agreement of the 3-ounce water swallow challenge has been established. A test value of 75 was identified as an acceptable metric (Hartmann, 1977; Stemler, 2004) and will serve as the comparative value for the pilot study. A paired sample *t*-test will compare the mean accuracy of scores across the cup and straw delivery methods. Cohen's Kappa coefficient (κ) will measure intrarater reliability of expert raters' initial and repeated CSV judgments. Fleiss' Multi-rater Kappa (Fleiss, 1971) will measure the level of interrater reliability.

Results

The pilot study yielded an 80% response rate and a 70% completion rate. There was a single partially completed rating task; the participant discontinued the online rating task during the reliability module. This rater's data was included in the interrater reliability calculation and excluded from the intrarater reliability calculation.

Accuracy

Overall Accuracy

A one-sample *t*-test was conducted to evaluate whether the expert SLP mean accuracy scores were significantly different from the test value of 75. Data from eight raters revealed a mean accuracy of 77% (Table 4). The lowest-performing rater achieved 64.7% accuracy. The highest-performing rater achieved an accuracy of 88%. 76.4% accuracy was the most frequently occurring score; this score was produced in five of eight opportunities. All other scores occurred a single time (Table 5).

Table 4

Percentage of Accuracy

N	Valid	8
	Missing	1
Mean		77.125
Median		76.400
Mode		76.4
Std. Deviation		6.5689
Minimum		64.7
Maximum		88.0

Table 5
Frequencies: Percentage of Accuracy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	64.7	1	11.1	12.5	12.5
	76.4	5	55.6	62.5	75.0
	82.3	1	11.1	12.5	87.5
	88.0	1	11.1	12.5	100.0
	Total	8	88.9	100.0	
Missing	System	1	11.1		
Total		9	100.0		

Expert mean accuracy (*Mexp* = 77.125, *SD* = 6.56) was not statistically different from 75, 95% CI[-3.36 to 7.61], $t(7) = .91$, $p = .39$ with a small effect size (d =.32; Cohen, 1988). These results indicate that there is no statistically significant difference between expert mean accuracy and the test value.

Accuracy According To Delivery Method

A paired-sample *t*-test was conducted to evaluate whether expert mean accuracy scores differed according to the delivery method (cup, straw). Expert mean cup accuracy (*M* = 80.17, *SD* = 7.48) was significantly greater than expert mean straw accuracy (*M* = 71.87, *SD* = 5.78), 95% CI [1.35 to 15.24], $t(7) = 2.82$, $p = .02$ with a large effect size (d =1.0; Cohen, 1988). This indicates that expert raters' CSV ratings were more accurate when rating the cup delivery method CSVs compared to straw delivery method CSVs.

Clinical Scenario Video (CSV) Analysis

Expert participant responses for each of the 17 CVSs were examined for frequency of the response consistent with the working standard (Table 6). Responses consistent with the working standard were observed in at least 90% of opportunities for 10 of 17 CSVs.

Table 6

Clinical Scenario Videos (CSVs): Frequency of Response Matching Working Standard

Video number		Response	$N = 8$	Frequency
Cup Delivery Method				
1.	Uninterrupted drinking	Pass	$n = 8$	100.0%
2.	Delayed cough	Fail	$n = 7$	87.5%
3.	Uninterrupted drinking	Pass	$n = 8$	100.0%
4.	Repeat instruction	Pass	$n = 4$	50.0%
5.	Laryngeal response	Pass	$n = 0$	0.0%
6.	Interrupted drinking & coughing	Fail	$n = 8$	100.0%
7.	Interrupted drinking & throat clearing	Fail	$n = 8$	100.0%
8.	Interrupted drinking, no cough nor throat clear	Fail	$n = 7$	87.5%
9.	Uninterrupted drinking with immediate cough	Fail	$n = 8$	100.0%
Straw Delivery Method				
10.	Uninterrupted drinking	Pass	$n = 4$	50.0%
11.	Repeat instruction	Pass	$n = 0$	0.0%
12.	Laryngeal response	Pass	$n = 0$	0.0%
13.	Interrupted drinking & coughing	Fail	$n = 8$	100.0%
14.	Interrupted drinking & throat clearing	Fail	$n = 8$	100.0%
15.	Interrupted drinking, no cough nor throat clear	Fail	$n = 8$	100.0%
16.	Uninterrupted drinking with immediate cough	Fail	$n = 8$	100.0%
17.	Uninterrupted drinking with immediate cough	Fail	$n = 8$	100.0%

The remaining seven CSVs demonstrated a varied frequency of response matching the working standard. The response matching the working standard for CSV two (delayed cough cup delivery method) and CSV eight (interrupted drinking cup delivery method) was identified by 84.5% ($n = 7$) of participants. The response matching the working standard for CSV four (repeated instruction cup delivery method) and CSV 10 (uninterrupted drinking straw method) was identified by 50% ($n = 4$) of participants. The response matching the working standard for CSV five (laryngeal response cup delivery method), CSV 11 (repeated instruction straw delivery method), and CSV 12 (laryngeal response straw delivery method) was identified by 0% ($n = 0$) of participants.

The above findings confirm there are statistically significant differences in CSV mean rating accuracy, which may be attributed to the delivery method (cup, straw).

Reliability

Intrarater Reliability

Intrarater reliability was assessed during the final instrument module. In this reliability module, participants were asked to rate the same 17 clinical scenario videos presented in randomized order. When examining intrarater reliability, it is reported in the literature that researchers may re-attempt a small portion of initial judgments following time delay (Suiter et al., 2014). The anonymous nature of this rating task prevented response tracking and respondent identification thereby prohibiting future reassessment. As such, all responses were obtained in a single encounter immediately following initial ratings. Participants were not made aware of this repetition.

Data from seven raters were included in intrarater reliability calculations. Crosstabulation of categories revealed an overall proportion of agreement range of 94.1-100%. As this does not account for chance agreement, Cohen's Kappa statistic is a superior metric to interpret the strength of agreement.

Cohen's Kappa was used to determine if each rater's individual judgments were consistent between the initial video rating and the reliability video rating. The use of Cohen's Kappa for statistical analysis requires that five assumptions are met: categorical variables are mutually exclusive of one another, judgments are paired, symmetric crosstabulation, rater judgments are made independent of other raters, and raters are "fixed." Each of the assumptions have been met in this case and Cohen's Kappa is an acceptable analysis.

Relevant results are presented below in Table 7. Four of seven raters demonstrated perfect strength of agreement as measured by a $\kappa = 1.0$. The remaining three raters demonstrated "very good" agreement as measured by Cohen's Kappa coefficients of $\kappa = .877$, $\kappa = .821$, and $\kappa = .821$, respectively. All raters demonstrated a statistically significant result ($p < .001$). Cohen's Kappa was statistically significantly different to zero (p=.000), rejecting the null hypothesis. This indicated that each expert raters' individual judgments were consistent between the initial video rating and the reliability video rating.

Table 7
Intrarater Reliability Summary

Rater	Overall proportion of agreement	Strength of agreement	Cohen's Kappa Coefficient
Rater 1	100%	Perfect	$\kappa = 1.0$, 95% CI [0.0, 0.0], $p < .001$.
Rater 2	94.1%	Very good	$\kappa = .877$, 95% CI [.664, 1.09], $p < .001$.
Rater 3	100%	Perfect	$\kappa = 1.0$, 95% CI [0.0, 0.0], $p < .001$.
Rater 4	94.1%	Very good	$\kappa = .821$, 95% CI [-0.164. .560], $p < .001$.
Rater 5	100%	Perfect	$\kappa = 1.0$, 95% CI [0.0, 0.0], $p < .001$.
Rater 6	n/a	n/a	n/a
Rater 7	100%	Perfect	$\kappa = 1.0$, 95% CI [0.0, 0.0], $p < .001$.
Rater 8	94.1%	Very good	$\kappa = .821$, 95% CI [.590, 1.126], $p < .001$.

Note. Table includes Case Processing Summary, First Rating * Reliability Rating Crosstabulation, and Symmetric Measures. To adjust for the limited CI as reported by asymptomatic standard error, a 95% CI manually calculated (asymptomatic standard error x 1.96).

Interrater Reliability

Fleiss' Kappa measures the beyond-chance agreement of more than two raters when variables utilize categorical data (McHugh, 2012). Use of Fleiss' Kappa coefficient requires the following assumptions: categorical variables are mutually exclusive of one another, judgments are measured via consistent categories, and rater judgments are made independent of other raters. It is important to acknowledge that Fleiss' Kappa also assumes that judgments are performed by non-unique raters. This pilot study and, presumably, the final study utilizes unique raters, as all raters will review and judge all the same clinical scenarios.

Data from eight raters were included in interrater reliability calculations. Fleiss' Multi-rater Kappa was performed to determine level of agreement between raters' interpretations of 17 3-ounce water swallow challenge clinical scenario videos. 12/17 videos revealed perfect agreement among raters, yielding a 70.5% overall proportion of agreement (Fleiss et al., 2003).

Because Fleiss' Kappa does not offer a scale to interpret strength, the use Cohen's Kappa scale has been suggested (Altman, 1999; Landis & Koch, 1977). There was 'good' agreement between raters and a statistically significant result, $\kappa = .649$, 95% CI [.570, .727], $p = .000$. Fleiss' Kappa was statistically significantly different to zero (p=.000), rejecting the null hypothesis. These results indicate good overall agreement between raters' interpretations of 17 3-ounce water swallow challenge clinical scenario videos.

Interrater Reliability by Delivery Method. Fleiss' Multi-rater Kappa was calculated for each rating category to determine the level of agreement between raters' 3-ounce water swallow challenge filtered by delivery method (cup, straw). Using Cohen's Kappa scale, the cup delivery method revealed "good" agreement between raters and a statistically significant result, $\kappa = .721$, 95% CI [.598, .845], $p = .000$ (table 8). The straw delivery method revealed "moderate" agreement between raters and a statistically significant result, $\kappa = .514$, 95% CI [.413, .616], $p = .000$ (table 9).

Fleiss' Kappa was statistically significantly different to zero (p=.000) for both delivery methods, rejecting the null hypothesis. The above findings confirm there are statistically significant differences in CSV rating agreement for both the cup and straw delivery methods, however, agreement for the cup delivery method was superior when compared to the straw delivery method.

Table 8

Cup Method Overall Agreement[a]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	**.721**	.063	11.453	**.000**	**.598**	**.845**

a. Sample data contains 9 effective subjects and 8 raters.

Table 9

Straw Method Overall Agreement[a]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	**.514**	.052	9.959	**.000**	**.413**	**.616**

a. Sample data contains 8 effective subjects and 8 raters.

Discussion

The purpose of the pilot study was to determine feasibility, formalize study questions, optimize clinical scenario videos (CSVs), and establish expert rater performance for comparison in the final investigation. This was accomplished by examining the accuracy and reliability of expert Speech-Language Pathologists across 17 novel and unique clinical scenarios and delivery methods.

Accuracy

Overall Accuracy

Results indicated no significant difference between expert Speech-Language Pathologists' mean accuracy rating scores and the test value of 75. When examining rater performance, it is noteworthy to mention that no single expert rater demonstrated all responses consistent with the working standard. The best performing rater, rater five, provided responses consistent with the working standard in all but two opportunities: CSV five laryngeal response cup delivery method and CSV 12 laryngeal response straw delivery method.

Accuracy According to Delivery Method

Results indicated a significant difference in expert Speech-Language Pathologists' cup and straw delivery method mean accuracy scores. Accuracy differed among delivery method with cup delivery method rating accuracy outperforming straw delivery method rating accuracy. Experts demonstrated higher mean accuracy when rating cup delivery method CSVs over straw delivery method CSVs.

The differences in rater performance between cup and straw delivery methods are likely multi-factorial, and cautious interpretation of the results is encouraged. The

investigator attributes the results to a combination of limited visibility and task complexity. Expert participants reported reduced stimuli visibility during straw delivery method CSVs. Expert participant feedback specific to the straw delivery method included the limited ability to visualize the cup contents and reliance on audio stimuli rather than visual stimuli. It is suspected that the complexity of the clinical scenario video may also contribute to the differences in accuracy across delivery methods.

Clinical Scenario Video (CSV) Analysis

10/17 CSVs revealed responses consistent with working standard by at least 90% of participants. The outstanding seven CSVs were further analyzed in an attempt to better understand the reduced frequency of responses matching the working standard.

CSV 2 (delayed cough) and CSV 8 (interrupted drinking) identified frequency of response matching the working standard was demonstrated by 87.5%. The investigator attributed this result to a single rater's incorrect interpretation of CSV 2 and CSV 8. Because this issue appears isolated to one rater and infrequently occurred, CSV 2 and CSV 8 will be included as final study items.

CSVs demonstrating the same behavior across different delivery methods were judged differently by expert participants. The "uninterrupted drinking" clinical scenario behavior (CSVs one and 10) was rated differently by expert participants. CVS one was completed with the frequency of response matching the working standard by 100% of participants, while CSV ten yielded only 50% of participant responses matching the working standard. The investigator attributes this to participant feedback reporting limited visualization of cup contents during straw delivery method. The investigator

recommends the elimination of video 10 from the final study because of limited visualization.

The “repeated instruction” clinical scenario behavior (CSVs four and 11) was rated differently by expert raters. CSV four revealed 50% of participant responses matched the working standard, while CSV 11 revealed 0% of participant responses matched the working standard. The observed variation between the same clinical scenario behaviors is likely multifactorial. The investigator suggests that CSV four’s results may suggest limited knowledge of protocol administration and interpretation guidelines. CSV 11 may be affected by both impaired knowledge of protocol interpretation and limited visualization of cup contents during the straw delivery method. As a result, CSV four will remain in the final study stimulus item, and CSV 11 be excluded from the final investigation.

CSVs demonstrating the “laryngeal response” clinical scenario (CSV five and 12) were the only CSVs that consistently produced 0% of responses matching the working standard. Of note, CSVs five and 12 obtained 100% interrater agreement, indicating that all raters assigned “the patient failed the water swallow challenge” judgment to these stimuli. The absence of responses matching the working standard and 100% interrater reliability was surprising to the investigator for several reasons. The investigator postulates several factors which may account for expert rater performance. First, it is possible that all respondents incorrectly judged this clinical scenario video as a failure due to limited awareness of the pass/fail interpretation criteria stipulated by the administration protocol. A second explanation for this variation may also be that all raters interpreted the standardized patient’s behavior as a genuine cough, which would result in

a true “fail” rating. Of note, the investigator coached the standardized patient to perform a throat-clear behavior during the recording of the clinical scenario videos. Distribution of this study item to a larger and more diverse participant sample may provide additional valuable data and supports item inclusion for the final investigation

Not all straw delivery method CSVs resulted in reduced reliability nor frequency of response matching the working standard. CSVs 12 through 17, which utilized the straw delivery method, revealed perfect interrater reliability and 100% of responses matching the working standard.

The remaining straw delivery method clinical scenario videos (videos 10, 11) produced significant variation. The investigator attributes these differences to reported difficulty with visualizing cup contents during the straw delivery method recommends that CSVs 10 and 11 are eliminated from the final study stimuli.

Reliability

Expert participants demonstrated “good” intrarater reliability and “good” overall agreement between raters’ 17 3-ounce WSC CSV judgments. When expert participant CSV ratings were stratified by delivery method, both the cup and straw delivery methods were statistically significant. However, expert rater judgments of CSVs presenting the cup delivery method outperformed the straw method. The significance of these reported differences should be interpreted cautiously, given participant feedback detailing the reduced visibility of straw delivery method CSVs. The investigator suspects that the variations reported are more likely related to video visibility rather than true differences in interpretation. To prevent the recurrence of this issue during the final study, the straw delivery method will be excluded

Limitations

This study is not without limitations. Although videos were randomized, it is possible that testing effects impacted reliability measures. It is also possible that respondents memorized or recorded responses from earlier rating task modules and later referenced personal notes during the reliability module. The impact of utilization of a trained standardized patient rather than a genuine patient is also unknown. Finally, this pilot study included a small sample size which may not be generalized to the greater population at large.

Conclusion

The pilot study confirmed expert Speech-Language Pathologists' ability to accurately and reliably interpret 3-ounce WSC CSVs across a variety of novel clinical scenarios and delivery methods. The current investigation revealed significant findings across CVSs and delivery methods. Preliminary findings are encouraging and provide support to further examine this concept during a final study. Refinement to final study questions and clinical scenario videos will be executed as presented in the discussion. Given the small sample size in the investigation, a larger and more heterogeneous participant sample is recommended.

Phase II: Final Investigation

A final study was conducted following the conclusion of the pilot study. Field testing performed with expert participants during the pilot investigation resulted in several changes to the phase II procedures. First, the CSV field was reduced from 17 to nine by the elimination of the straw delivery method. The CSV field reduction was motivated by expert reports of poor visibility during pilot testing. As a result, the expert mean of *Mexp* = 80.47, *SD* = 7.84 is presented as the updated comparative value in the final investigation. This new comparative value corresponded to the expert mean accuracy obtained when only CSVs one through nine are examined. Next, participant training was revised to include knowledge of performance with feedback. Finally, YSP interpretation guidelines were made available on demand.

The final study was reviewed by the University IRB (Influence of Rater Experience on Water Swallow Testing Interpretation Protocol 1838584-2, exempt).

The investigator proposed three research questions to be addressed during the final study.

Research Questions

Research Question 1. Is the accuracy of judging 3-ounce water swallow challenge (WSC) clinical scenario videos (CSVs) from non-expert SLPs the same as expert SLPs?

Null hypothesis: There is no significant difference in non-expert mean accuracy and expert mean accuracy (80.47). H0: $\mu_{non} = 80.47$

Alternate hypothesis: Non-expert mean accuracy and expert mean accuracy are not equal. HA: $\mu_{non} \neq 80.47$.

Note: CSVs illustrated a standardized patient trained to perform the 3-ounce WSC portion of the Yale Swallow Protocol. A test value of 80.47 was identified as the expert mean accuracy during pilot testing.

Research Question 2. To what extent can non-expert SLPs reliably judge 3-ounce water swallow challenge (WSC) videos?

Null hypothesis: Agreement between and within rater responses is no different than chance agreement. H0: $\kappa = 0$.

Alternative hypothesis: Kappa (κ) coefficient is different from the zero/chance agreement. HA: $\kappa \neq 0$

Research Question 3. Can 3-ounce water swallow challenge (WSC) rating accuracy be predicted from a combination of non-expert SLPs' demographic information?

Null hypothesis: All regression coefficients are equal to zero.

Alternate hypothesis: Some of the regression coefficients are not equal to zero.

Procedure

Participants and Recruitment/Sampling Procedure

SLPs were invited to participate in the study examining clinical decision-making. All levels of experience and clinical settings were eligible for study participation. An invitation for study participation was presented among multiple social media platforms (Facebook.com, Instagram.com, LinkedIn.com), ASHA electronic listservs (Special Interest Group 3, Special Interest Group 13, ASHA members, Autism, Clinicians & Researchers Collaborating, Early Intervention, Professional Materials Exchange, Research, Rural and Remote Service Delivery, SLP Health Care, SLP Private Practice, SLP Schools). The investigator also shared the study invitation with SLPs via direct

invitation. As the electronic link to participate in the study was non-unique to the user and shareable, SLPs may have been recruited to participate via referred sampling. Study recruitment and data collection were active for 14 days (May 28, 2022, through June 11, 2022).

Inclusion criteria and study participation was limited to ASHA certified SLPs holding the Certificate of Clinical Competence (CCC) and clinical fellows currently residing in the United States. Participants who received their education/training in Canada before their United States residency were eligible for participation. Participants were excluded if they self-reported previous completion of the study. There were no population criteria for gender and ethnic background.

Instrument

The investigator generated 9 novel and unique clinical scenarios to be used as research stimuli. The clinical scenarios represented by the CSVs were created from the investigator's own clinical experiences and observations during YSP administration. CSVs were created by the investigator to allow for the examination of task complexity and YSP interpretation. Of note, the YSP does not delineate protocol interpretation beyond pass-fail criteria.

A University of Nevada, Reno School of Medicine standardized patient was hired to portray a patient demonstrating a 3-ounce WSC. The investigator trained the standardized patient to perform a 3-ounce WSC for each clinical scenario represented in Table 10 below.

It should be noted that the nine CSVs presented in the current investigation were selected from a larger field of 17 CSVs presented during the pilot investigation. Consult

phase I discussion section for supplemental information. Standardized patient performance was video recorded and edited by the investigator. Raw videos were trimmed to eliminate excess video before and after the 3-ounce WSC. Additionally, different video image angles (anterior-posterior and lateral) were combined into a single side-by-side video image. Finalized videos were uploaded to the investigator's unlisted YouTube channel (Morrissey, n.d.), converted into HTML format, and embedded into Qualtrics Experience Management Software (Qualtrics, Provo, UT) survey template.

Table 10
Clinical Scenario Videos (CSVs) Presented for Non-Expert Rater Interpretation

Video	Working standard response	Clinical scenario behavior
1.	Pass	Uninterrupted drinking
2.	Fail	Delayed cough
3.	Pass	Uninterrupted drinking
4.	Pass	Repeat instruction
5.	Pass	Laryngeal response (throat clear vs cough)
6.	Fail	Interrupted drinking & coughing
7.	Fail	Interrupted drinking & throat clearing
8.	Fail	Interrupted drinking, no cough nor throat clear
9.	Fail	Uninterrupted drinking with immediate cough

Instrument Modules. Based upon evidence for standardized clinical training and consensus, the instrument included three modules: learning, training, and demographics. The learning and training modules were modeled upon the Modified Barium Swallow Impairment Profile (MBSImP; Martin-Harris et al., 2008), a standardized protocol for interpreting video fluoroscopic swallow studies. The three-module format was modeled upon the Modified Barium Swallow Impairment Profile (MBSImP; Martin-Harris et al., 2008), a standardized protocol for interpreting video fluoroscopic swallow studies. Martin-Harris reports that the MBSImP three-module design was self-generated using a

"logical approach logical approach to other competency-based learning in our clinics" (personal communication, June 22, 2022). Although Martin-Harris did not follow adhere to a particular theoretical model during MBSImP generation, the integration of psychometrically sound training practices (training practice with feedback, and competency assessment) are readily apparent.

The Learning Module. presented a 2-minute video tutorial providing participant training on the administration and interpretation of the 3-ounce water swallow challenge task of the YSP. The video tutorial included verbal instruction and visual supports detailing the pass/fail criteria and YSP definitions. Participants received the following instruction:

> "The Yale Swallow Protocol is a validated screening method designed to reliably and accurately identify aspiration risk. In this video, you will learn to administer and interpret a single component of the Yale Swallow Protocol; the 3-ounce WSC. It's important to note that the Yale Swallow Protocol is one of several water swallow tests. Each test presents its own administration and interpretation guidelines. For the purposes of this study, we are looking only at the Yale Swallow Protocol. The Yale Swallow Protocol has three components: a cognitive screening, oral mechanism exam, and the 3-ounce WSC. The patient is provided a cup containing three ounces of water and provided the following instructions "drink this water, slow and steady, without stopping." Performance is interpreted by observing behavior during and immediately following the conclusion of drinking. The Yale Swallow Protocol is a screening measure and interpretation is stratified into pass/fail results. According to the flow diagram presented in the

> Yale Swallow Protocol, pass criteria includes complete and uninterrupted drinking of all 3 oz of water and with no overt signs of aspiration during or immediately after completion. Fail criteria is described as interrupted drinking, coughing, or choking during or immediately after completion of drinking. It is important to highlight that in Leder and Suiter's (2014) description, overt signs of aspiration are defined only by coughing or choking. As mentioned earlier, there are several other water swallow tests each with their own interpretation and administration guidelines. For the purpose of this study, we are considering only the Yale Swallow Protocol. Let's practice."

After reviewing the 2-minute instructional video, participants were presented with two practice CSVs. Participants reviewed each practice CSV and were prompted to apply the information presented in the instructional video tutorial to interpret the Yale Swallow Protocol's 3-ounce WSC. Participants evaluated standardized patient performance by applying one of three possible responses: *"The patient passed the water swallow challenge," "The patient failed the water swallow challenge,"* and *"I'm not sure."* Participants were provided with an "*I'm not sure*" response to avoid forcing of a response when the response was not known (Vose et al., 2018). Participants were permitted unlimited review of the instructional video(Stoeckli et al., 2003). Participants received knowledge of their performance and were given immediate feedback. Once answered, participants were not allowed to return to a previously answered CSV. Participants then transitioned to the training module.

The Training Module. required that participants apply the knowledge obtained during the instructional video. The training module presented participants with 11 CSVs (9 novel + 2 repeated) for the purposes of obtaining accuracy and reliability data.

Participants were presented with each CSV and were prompted to apply the information presented in the instruction video tutorial to interpret the YSP 3-ounce WCS.

Participants evaluated standardized patient performance by applying one of three possible responses: *"The patient passed the water swallow challenge," "The patient failed the water swallow challenge,"* and *"I'm not sure."* Participants were provided with an "*I'm not sure*" response to avoid forcing of a response when the response was not known (Vose et al., 2018). Definitions were available on demand. Participants were permitted unlimited review of the CSV (Stoeckli et al., 2003). Participants did not receive feedback on their performance. Once answered, participants were not allowed to return to a previously answered CSV. Participants then transitioned to the demographic module.

The Demographic Module. presented eight questions including dysphagia-related years of clinical experience, primary practice setting, weekly dysphagia-related care in hours providing, highest level of education, professional certifications and skills, academic coursework, and use of water swallow screenings. Demographic questions presented in the current study were adopted from Vose and colleagues' (2018) survey investigating SLP demographics as they relate to the identification of swallowing impairments and treatment recommendations.

Operational Definitions and Scoring

The scoring of participants' CSVs judgments integrated the use of a reference standard and a working standard. According to Cook (2012), "a reference standard refers

to the best available method for establishing the presence or absence of a condition of interest" (p. 111). A reference standard functions as the standard by which judgments are evaluated. Leder and Suiter's (2014) Yale Swallow Protocol interpretation definitions were selected by the investigator as the reference or gold standard following the proven sensitivity and specificity to accurately identify aspiration.

The investigator applied Leder and Suiter's (2014) reference standard to the novel CSVs presented in the pilot and final studies. This generated a "working standard" used for comparing participant CSV ratings and scoring for accuracy. A working standard is a measurement standard used for analysis and standardized against the reference standard (Ph. Eur., 2015). For the purpose of the study, the working standard is calibrated against a reference standard and is used to measure participant adherence to the reference standard (YSP). Participant responses were qualified by a working standard (AM) and scored for accuracy (Martin-Harris et al., 2008).

A working standard was applied for both expert and non-expert SLPs with consistent application of definitions and scoring during the pilot and final investigations and scored for accuracy.

Accuracy in this study is defined as participant ratings identical to those of the working standard as an established criterion. Participants were credited with correctly responding when their ratings matched the working standard. Accurate ratings received a "1" whereas inaccurate ratings received a "0." The participant score was computed by calculating the number of accurate responses divided by the total number of CSVs (nine).

In this study, reliability over time and between raters is defined as the consistency of CSV judgments. The scoring procedures for both accuracy and reliability were completed in the same manner for each group.

When examining intrarater reliability, it is reported in the literature that researchers may re-attempt a small portion of initial judgments following time delay (Suiter et al., 2014). The anonymous nature of the study prevented response tracking and respondent identification, prohibiting future reassessment. All responses were obtained in a single encounter immediately following initial ratings. Intrarater reliability was assessed by an at-random repeat presentation of two (20%) previously viewed CSVs. Participants were not made aware of the repetition.

Data Collection*:* 9 metadata variables including start date, end date, response type, progress, duration, completion, recorded date, distribution, user language (excluded from data extraction), 2 practice videos (excluded from data extraction), participant score, 11 videos, and eight demographic questions.

Data Extraction: 11 CSVs (9 novel plus 2 repeated), 8 demographic questions.

- Nine CSVs comparing the accuracy of participant ratings to the working standard
- Two repeated CSVs comparing the consistency of one rater's responses across two separate rating opportunities
- Eight questions using participant attributes as a predictor for response accuracy.

Data Analysis: Participant responses were maintained by Qualtrics Experience Management Software (Qualtrics, Provo, UT). All responses are anonymous and no identifying information from the rater is associated with rating task responses. Rater

responses cannot be traced to the participant. Anonymous Qualtrics data was extracted, coded by the investigator, and imported into SPSS Statistics for Windows (Version 28.0.1) for analysis.

Statistical Analysis

An *A Priori* G*Power Statistical Power Analysis (Faul et al., 2009) for a multiple regression model determined a medium effect size (0.15), alpha level.05, and power 0.80 required a total sample size of 109 participants.

A One-sample t-test will compare the mean accuracy of non-expert SLP ratings to experts. An expert mean accuracy of 80.47% was identified during pilot testing and will be used as the comparative value. Cohen's Kappa coefficient will measure intrarater reliability. Fleiss' Multi-rater Kappa (Fleiss, 1971) will measure the level of interrater reliability. Descriptive statistics will be used to report demographic frequencies, parentages, and means. A binomial regression will determine how non-expert SLP demographics predicted mean CSV rating accuracy. The outcome variable will be condensed from percent accuracy and stratified into a dichotomous variable based on accuracy scores above or below expert mean accuracy (M=80.47, SD =7.84).

Results

151 SLPs completed the online rating task. One participant's data was considered an extreme outlier and was removed from the analysis. 150 participants were included in the data analysis. Descriptive statistics are reported in the demographic section below.

Accuracy

A one-sample *t*-test was conducted to evaluate whether participant mean accuracy (M_{non} = 90.06) was significantly different from expert accuracy (M_{exp} = 80.47).

Boxplot inspection identified two outliers and one extreme score. The single extreme score (22% accuracy) was considered an unusual value and was removed from the sample. The two outlier scores remained in the sample.

The skewness of participant mean accuracy was -1.01, indicating that the distribution was left-skewed. The kurtosis of participant mean accuracy was 1.90, suggesting that the distribution was heavily peaked compared to the normal distribution. A Shapiro-Wilk's Test was statistically significant ($W = .78$, $p < .001$), which implied the violation of the assumption of a normal distribution. Due to the violation of the normality assumption, the non-parametric counterpart of one-sample *t*-test was performed. A Wilcoxon signed-rank test was performed to compare the median rating task accuracy scores.

Cohen's (1988) effect size for standardized mean difference (*d*) was computed using values 0.2, 0.5, and 0.8, which correspond to small, medium, and large effect sizes, respectively. Combined, these analyses determined there were significant differences in non-expert SLP mean and median rating accuracy scores when compared to the scores of expert SLPs. Non-expert SLP mean accuracy (*Mnon* = 90.06, *SD* = 8.45) was

significantly higher than expert mean accuracy (*Mexp* = 80.47, *SD* = 7.84) by a mean difference of 9.59, 95% CI [8.23 to 10.95], $t(149) = 13.89$, $p < .0005$ with a large effect size ($d = 1.13$; Cohen, 1988). Non-expert SLPs demonstrated a statistically significant median increase in rating task accuracy scores, $z = 10.53$, $p < .0005$. This confirmed that non-expert SLPs outperformed expert SLPs with greater mean and median rating task accuracy scores, rejecting the null hypothesis

Clinical Scenario Video (CSV) Analysis

Non-expert participant responses for each of the nine CSVs were examined for frequency of the response consistent with the working standard (Table 11). Responses consistent with the working standard were observed in at least 90% of opportunities for seven of the nine CSVs. Two CSVs (four and five) revealed significantly reduced frequency of the response matching the working standard. The response matching the working standard for CSV four (repeat instruction) was identified by 42% (n=63) of participants. The response matching the working standard for CSV 5 (laryngeal response) was identified by 11.3% (n=17) of participants.

Table 11
Clinical Scenario Videos (CSVs): Frequency of Response Matching Working Standard

CSV		Response	$N = 150$	Frequency
1.	Uninterrupted drinking	Pass	$n = 150$	100.0%
2.	Delayed cough	Fail	$n = 147$	98.0%
3.	Uninterrupted drinking	Pass	$n = 148$	98.7%
4.	**Repeat instruction**	**Pass**	***n* = 63**	**42.0%**
5.	**Laryngeal response (throat clear)**	**Pass**	***n* = 17**	**11.3%**
6.	Interrupted drinking & coughing	Fail	$n = 150$	100.0%
7.	Interrupted drinking & throat clearing	Fail	$n = 147$	98.0%
8.	Interrupted drinking, no cough nor throat clear	Fail	$n = 135$	90.0%
9.	Uninterrupted drinking with immediate cough	Fail	$n = 149$	99.3%

Reliability

Intrarater Reliability

Intrarater reliability was assessed by an at-random repeat presentation of 20% (two) previously viewed CSVs. All responses were obtained in a single encounter immediately following initial ratings. Participants were not made aware of the repetition.

Cohen's Kappa was used to determine the consistency of rater's initial and repeated CSVs judgments. When attempted, a crosstabulation of categories revealed Kappa as an invalid statistic. This limitation, attributed to high levels of agreement and a small sample size, has been described as a limitation to the Kappa statistic (Zec et al., 2017). Additional statistical analyses were limited by the categorical nature of the study variables.

The overall proportion of agreement is reported in when the Kappa statistic cannot be calculated. This was accomplished by comparing participants' ratings on two (20%) repeated CVSs. The initial rating and the reliability ratings were compared for agreement. The participant received a score of "1" when initial and repeated ratings were the same. The participant was assigned a score of "0" if the initial and repeated ratings were not the same. Scores of "0" indicated inconsistent participant ratings for both repeated CSVs. Scores of "50%" indicated consistent participant ratings in one of two repeated CSVs. Scores of "100%" indicated consistent participant ratings in two of two repeated CSVs.

Informal indices have reported expected intrarater agreement at 90% or greater (Roache, 2017). The overall proportion of agreement for each individual rater ranged from 50-100%. 91% of participants (n = 137) demonstrated exact agreement (100%)

between their first and second CSV ratings. 8.7% of participants (n = 13) demonstrated partial agreement (50%) between their first and second CSV ratings. All participants demonstrated some level of agreement between their first and second CSV ratings; absent agreement (0%) between first and second CSV ratings did not occur. See table 12 below.

Table 12
Intrarater Reliability (Per Rater) Between First and Second CSV Ratings

Overall Proportion of Agreement	Frequency	Percentage
0% (No ratings in agreement)	n = 0	0
50% (One of two ratings in agreement)	n = 13	8.7
100% (Two of two ratings in agreement)	n = 137	91.3

Note. N=150.

Interrater Reliability

Fleiss' Multi-rater Kappa was performed to determine the level of overall agreement of between raters' interpretations of the nine combined 3-ounce WSC CSVs. Fleiss' (1971) Kappa measures the beyond chance agreement of more than two raters when variables utilize categorical data (McHugh, 2012). Using Fleiss' Kappa coefficient requires the following assumptions: categorical variables are mutually exclusive of one another, judgments are measured via consistent categories, and rater judgments are made independent of other raters. It is essential to acknowledge that Fleiss' Kappa also assumes that non-unique raters perform judgments. The current investigation utilizes unique raters, as all raters will review and judge the same CSVs.

Because Fleiss' Kappa does not offer a scale to interpret the strength of agreement, the use of Cohen's Kappa scale has been suggested (Altman, 1999; Landis & Koch 1977). Strength of agreement for Cohen's Kappa included the following thresholds: poor ($\kappa \leq .20$), fair ($\kappa = .21$-$.40$), moderate ($\kappa = .41$-$.60$), good ($\kappa = .61$-$.80$), and very good ($\kappa = .81$-1.00; Altman, 1999).

Kappa was significant for "good" agreement of the nine combined CSVs (κ = .69, 95% CI [.692, .703], $p < .001$). Fleiss' Kappa was significantly different to zero (p=.000), rejecting the null hypothesis and failing to reject the alternative hypothesis. This indicates that non-expert SLPs are consistent in their CSV ratings.

Demographics

Research question three focused on the demographic characteristics of non-expert SLP participants (see tables 13 and 14). One hundred fifty-one respondents participated in this study. 92% (n = 139) of participants reported an ASHA Certificate of Clinical Competence with an average of 11 years (M=11.23, SD=10.75) of experience. Participants reported an average of 15 hours (M=15.53, SD=13.56) of weekly dysphagia care (evaluation, screening, intervention). 88% (n = 132) of participants were Master's level clinicians with only 1.3% (n = 2) pursuing or 10.7% (n=16) having completed a terminal degree. 53% (n=80) of participants are employed at a hospital for their primary work setting. 76% (n=114) of participants endorsed the completion of a dedicated swallowing course during their academic training.

Regarding professional certifications and skills, 64.7% (n=97) of participants endorsed using Water Swallow Tests (WSTs) in their clinical practice. Of those that included WSTs in their clinical practices, daily (25.3%, n=38) and weekly (21.3%, n=32) use were the most common frequencies. 34.7% (n=52) reported Modified Barium Swallow Impairment Profile certification. 3.3% (n=5) identified as a Board-Certified Specialist in Swallow and Swallowing Disorders. 18% (n=27) endorsed McNeill Dysphagia Therapy Program certification. 22% (n=33) of respondents endorsed VitalStim certification, compared to the 10% (n=15) endorsed Ampcare ESP training.

24% *(n*=37) were Lee Silverman Voice Therapy providers and 12% (*n*=18) were SPEAK OUT! providers. 6% (*n*=9) of participants endorsed deep pharyngeal neuromuscular stimulation. 40% (*n* = 60) endorsed training in flexible endoscopic evaluation of swallowing, while 57% (*n* = 86) identified training/skills in modified barium swallow studies. 2.7% (*n*=4) of participants endorsed training or skills in high resolution pharyngeal manometry.

Table 13
Summary of Rater Demographic Responses

Certification & Skills		Frequency	Percentage
CCC-SLP	No	11	7.3
	Yes	139	92.7
Years of dysphagia experience	*M* = 11.23, *SD* = 10.75	n/a	n/a
Dysphagia hours weekly	*M* = 15.53, *SD* = 13.56	n/a	n/a
Education	Master's degree	132	88.0
	terminal degree in progress	2	1.3
	terminal degree completed	16	10.7
Primary practice setting	Hospital	80	53.3
	Residential healthcare facility	16	10.7
	Non-residential healthcare facility	24	16.0
	Community dwelling	9	6.0
	School	19	12.7
	Tele-health	2	1.3
Swallow Course	No	36	24.0
	Yes	114	76.0
WST use	Yes	97	64.7
	No	53	35.3
WST frequency	Does not use	53	35.3
	Daily	38	25.3
	Weekly	32	21.3
	Monthly	17	11.3
	Annually	10	6.7
MBSImP	No	98	65.3
	Yes	52	34.7
BCSS	No	145	96.7

	Yes	5	3.3
MDTP	No	123	82.0
	Yes	27	18.0
VitalStim	No	117	78.0
	Yes	33	22.0
Ampcare ESP	No	135	90.0
	Yes	15	10.0
LSVT	No	113	75.3
	Yes	37	24.7
SPEAK OUT!	No	132	88.0
	Yes	18	12.0
DPNS	No	141	94.0
	Yes	9	6.0
FEES	No	90	60.0
	Yes	60	40.0
MBSS	No	64	42.7
	Yes	86	57.3
HRPM	No	146	97.3
	Yes	4	2.7

Note. CCC-SLP = Certificate of Clinical Competence in Speech-Language Pathology; WST = water swallow test; MBSImP = Modified Barium Swallow Impairment Profile; BCSS = Board-Certified Specialist in Swallow and Swallowing Disorders; MDTP = McNeill Dysphagia Therapy Program; LSVT = Lee Silverman Voice Therapy; DPNS = deep pharyngeal neuromuscular stimulation; FEES = flexible endoscopic evaluation of swallowing; MBSS = modified barium swallow study; HRPM = high resolution pharyngeal manometry.

Binomial logistic regression/multiple logistic regression analyzed the relationship and prediction ability of the independent/predictor variables (demographic data) on the dependent/outcome variable (rating accuracy score). This analysis was selected to predict whether non-expert participants rating task accuracy scored above or below the expert mean accuracy based on a combination of demographic characteristics. In binomial regression, the categorical variables interpret the odds that one group has higher or lower

accuracy. Continuous variables interpret how a single unit change (increase or decrease) is associated with the odds of a change in accuracy (increase or decrease).

This form of analysis requires seven assumptions; four assumptions related to study design and three assumptions related to model fit. The first assumption requires a dichotomous dependent or outcome variable. Next, the independent or predictor variables must be measured on a nominal scale or as a continuous variable. The third assumption requires an independence of observations. The fourth assumption stipulates that there is a minimum of 10-20 events per covariate (Stoltzfus, 2001). Assumption five requires linearity of the continuous independent variables (years of experience and dysphagia hours weekly). Assumption six required that independent variables are not correlated. Assumption seven instructs that there should be no significant outliers.

The 18 demographic predictor variables included years of experience, dysphagia hours weekly, CCC-SLP status, MBSIMP status, BCSS status, MDTP status, VitalStim status, AmpCare status, LSVT status, SpeakOut status, DPNS status, FEES status, MBSS status, HRPM status, history of a dedicated swallow course, WST frequency, education level, and practice setting.

Table 14
Clinical Characteristics of Participants

Demographic stimuli	Response	Frequency	Percentage
Water swallow screening frequency	I don't use this (R).	53	35.3%
	I use this daily.	38	25.3%
	I use this more than weekly but not daily.	32	21.3%
	I use this monthly but not weekly.	17	11.3%
	I use this a few times per year.	10	6.6%
Terminal degree completed	None (R).	134	89.3%
	Terminal degree completed.	16	10.6%
Dedicated swallowing course	No (R).	36	24%
	Yes.	114	76%
I use this daily	None (R).	112	74.6%
	I use this daily	38	25.3%
More than weekly but not daily	None (R).	118	78.6%
	I use this weekly but not daily	32	21.2%
Monthly but not weekly.	None (R).	133	88.6%
	Monthly but not weekly	17	11.3%
Few times per year.	None (R).	140	93.3%
	A few times a year	10	6.6%
Hospital setting	None (R).	70	46.6%
	Hospital setting	80	53.3%
Residential Healthcare Facility	None (R).	134	89.3%
	Residential healthcare facility	16	10.6%
Terminal degree in progress	None (R).	148	98.6%
	Terminal degree in progress	2	1.3%
Master's degree completed.	None (R).	18	12%
	Master's degree completed	132	88%
Tele-health setting	None (R).	148	98.6%
	Tele-health setting	2	1.3%

School setting	None (R).	131	87.3%
	School setting	19	12.6%
Non-Residential Healthcare Setting	None (R).	126	84%
	Non-residential healthcare setting	24	16%
Community dwelling	None (R).	141	.000
	Community dwelling	9	1.000
Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)	None (R).	11	.000
	Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)	139	1.000

Note. For each variable, the category with (R) was coding as the reference category (= 0).

The outcome variable of percent accuracy was condensed from five categories (55.5%, 66.6%, 77.7%, 88.8%, 100.0%) and stratified into a dichotomous variable based on accuracy scores above or below expert mean accuracy (M=80.47, SD =7.84). Non-experts were classified as either scoring above the expert mean (n=129, 86% frequency) or below the expert mean (n=21, 14% frequency).

Years of experience and dysphagia hours weekly required the creation of natural log transformation with interaction terms. The Box-Tidwell Procedure (1962) was conducted to evaluate the linearity of the continuous independent variables. As reflected below, the interaction terms are not significant, which indicated that the original continuous independent variables (dysphagia experience in years and weekly dysphagia care hours) were linearly related to the logit of the dependent variable.

Clinical skills and certification demographics violated the assumption of multicollinearity. This indicated a high correlation between the various levels or categories of the demographic questions. This failed the assumptions for parametric tests, and non-parametric statistical analysis was completed. The aforementioned variables

violated the assumption of multicollinearity and were removed from the regression analysis. As a result, the demographic predictor variables were reduced from 18 to seven variables: years of experience, dysphagia hours weekly, CCC-SLP status, history of a dedicated swallow course, WST frequency, education level, practice setting.

Somers' delta (*d)* was utilized to determine the strength and direction of the association between the dependent variable and individual clinical skills/certifications (Somers, 1962). Neither CCC-SLP status, MBSImP status, MDTP status, VitalStim status, Ampcare ESP status, LSVT status, Speak OUT! status, DPNS status, MBS status, BCSS status, nor FEES status were statistically significant.

The Casewise list identified eight cases with standardized residuals greater than 2.5 standard deviations. The cases remained in the analysis.

There were conflicting statistics regarding model fit. An Omnibus Test of Model Coefficients was performed to determine model significance and prediction (Table 15). The Omnibus Test of Model Coefficients determined that the model was not statistically significant $\chi^2(15) = 6.55$, $p = .97$. This indicated that when compared to the model with no independent variables, this model containing the demographic variables was a poor predictor of non-expert participant rating accuracy category (above expert mean, below expert mean).

Table 15

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	6.554	15	.969
	Block	6.554	15	.969
	Model	6.554	15	.969

A Hosmer and Lemeshow goodness of fit test indicated the model was not a poor fit and was not significant (p = .310). The model summary identified the variation in the outcome variable of the accuracy rating category (above expert mean, below expert mean) was attributed to the model. The explained variation in the dependent variable based on the model ranged from a Cox & Snell R square .04 or 4.3% to the Nagelkerke R square value .07 or 7.7%. (Table 16).

Table 16

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	114.935[a]	.043	.077

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Percentage accuracy in classification (PAC; Table 17) revealed that the model correctly predicted 87% of cases overall. This prediction can be compared to the original model without independent variables that correctly predicted 86% of cases. This indicated that the addition of the predictor variables to the model increased overall prediction of cases by 1%. The model accurately predicted 100% of cases scoring above the expert mean accuracy (sensitivity). 5% of cases that scored below expert mean accuracy were accurately predicted by the model (specificity).

The area under the ROC curve (figure 3) was 0.58, 95% confidence interval CI [0.45, 0.70] which corresponded to a “poor” level of discrimination (Hosmer et al., 2013). The variables presented in the equation are illustrated below. Of the predictor variables included in the model, none of the predictor variables were statistically significant. This indicated that the non-expert SLP rating task accuracy rating category could not be predicted by a combination of demographic characteristics.

Table 17

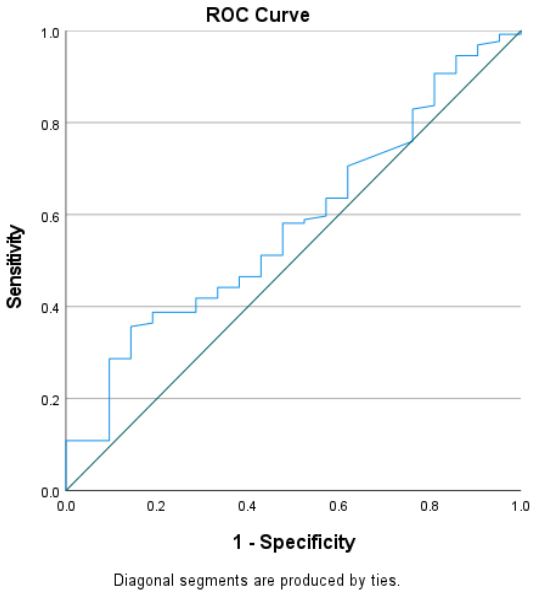
Classification Table[a]

			Predicted		
			Percent accuracy compared to experts		
	Observed		Below expert mean	Above expert mean	Percentage Correct
Step 1	Percent accuracy compared to experts	Below expert mean	1	20	4.8
		Above expert mean	0	129	100.0
	Overall Percentage				86.7

a. The cut value is .500

Figure 3

ROC Curve



Discussion

The purpose of the current study sought to investigate the swallow screening practices of Speech-Language Pathologists. This was accomplished in two ways. First, the investigator examined the accuracy and reliability of participant responses across nine unique and novel clinical videos. Second, the investigator explored the relationship between Speech-Language pathologists' clinical video ratings and their respective demographic features.

Accuracy

Results indicated a significant difference between the median scores of participant accuracy and expert accuracy. The study results confirmed that non-expert SLP participants demonstrated higher mean and median accuracy compared to expert raters. Higher mean and median accuracy observed in the non-expert SLP group was thought to be attributed to the availability of on-demand definitions and training with feedback on performance.

Clinical Scenario Video (CSV) Analysis

Participant responses for each of the nine CSVs were examined for frequency of the response consistent with the working standard. Responses consistent with the working standard were observed in at least 90% of opportunities for seven of the nine CSVs.

Two CSVs (four and five), however, revealed significantly reduced frequency of response matching the working standard. The investigator attributes the differences observed in the CSV judgments to the complexity of the clinical behavior represented on the respective CSVs. As showing in the previous literature review, the complexity of

stimuli has been shown to impact rater judgments (Ekberg et al., 1988; Martin-Harris, 2015; Scott Perry & Bench, 1998; Vose et al., 2018).

CSV four presented the clinical scenario of repetition of instructions and warranted a "pass" judgment per the YSP. Recall that YSP defines "pass" criteria as "complete and uninterrupted drinking of all 3 oz of water and with no overt signs of aspiration during or immediately after completion" (Leder & Suiter, 2014). Of note, neither the instructional video presented in module one nor the on-demand definitions addressed this scenario, should it occur. Therefore, the investigator hypothesized only raters familiar with the full YSP text, not simply the pass/fail criteria, would likely produce a response matching the working standard.

When examining the demographic data, 65% of non-expert participants endorsed WST use in their clinical practice. Of the 65% reporting WST use, 55% of those non-expert participants assigned the inaccurate "fail" judgment. Recall that a "fail" judgment should be applied for "interrupted drinking, coughing, or choking during or immediately after completion of drinking" (Leder & Suiter, 2014).

Although a repetition of instructions is permitted by the YSP, the investigator attributes the low frequency of response matching the working standard observed in CSV four to poor YSP knowledge. This variation in performance was not completely unexpected. When presented during phase I pilot testing with expert raters, CSV four performed only slightly worse when it yielded a frequency of response matching the working standard in 50% of opportunities.

These findings suggest that nearly half of Speech-Language Pathologists using WST failed to appropriately interpret the YSP. With the support of the existing body of

literature (Clain et al., 2021; Hind et al., 2009; Martin-Harris et al., 2008; Scott, Perry,& Bench, 1998), the investigator asserts that formalized training would likely optimize the clinical utility of the YSP.

CSV five presented the clinical scenario of laryngeal response and warranted a "pass" judgment per the YSP 3-ounce WSC. Recall that YSP defines "pass" criteria as "complete and uninterrupted drinking of all 3 oz of water and with no overt signs of aspiration during or immediately after completion" (Leder & Suiter, 2014). The instructional video presented in module one instructed raters that "overt signs of aspiration" were defined as "coughing or choking during or immediately following the conclusion of drinking" (Leder & Suiter, 2014). Therefore, the investigator hypothesized that raters adherent to the YSP interpretation guidelines presented during module one would likely produce a response matching the working standard.

The frequency of response matching the working standard was produced in 11.9% of opportunities. The very low frequency of response matching the working standard was thought to be multifactorial. One consideration for the low frequency of response matching the working standard is that participants judged this CSV as a failure due to limited awareness of the pass/fail interpretation criteria stipulated by the administration protocol (i.e., any laryngeal response results in failure). Because the YSP is just *one* of many WSTs, it is possible the participants erroneously applied the interpretation guidelines belonging to a separate WST despite instruction to apply only YSP definitions for interpretation.

An additional explanation for this variation may include that raters interpreted the standardized patient's laryngeal response as a "cough" rather than the intended "throat

clear." The standardized patient was coached by the investigator to perform a "throat clear," which aligns with YSP "pass" judgment; however, this clinical scenario may have been interpreted by the rater as a "cough." A "cough" response would yield a "fail" judgment per YSP interpretation guidelines.

A final consideration related to CSV five is the absence of definitions. It is reasonable that a rater might inaccurately apply the protocol interpretation guidelines when definitions to "coughing" and/or "choking" are not provided. Prior research has confirmed the positive impact of clear definitions on rater reliability (Stoeckli et al., 2003). In the absence of formalized training with definitions or anchors, it is possible that participants formulated their own internal standards rather than adhering to the protocol or external standards.

CSV five identified that 87% of Speech-Language Pathologists using WST assigned a "fail" judgment. The variation in performance compared to the working standard was not completely unexpected. Of note, the expert rater frequency of response matching the reference standard for CSV five was produced in 0% of opportunities. This indicates that all of expert raters assigned CSV five a "fail" response during pilot testing. If expert rater knowledge of the YSP interpretation guidelines is assumed, expert rater performance highlights the subjectivity with which one may interpret a laryngeal response ("cough" versus "throat clear"). While this was not the goal of the current research, future research may include participant judgments and the rationale when assigning clinical judgment to clinical scenarios.

Reliability

Interrater reliability examined the consistency of participant ratings across the nine CVSs. Results indicated "good" agreement among both expert and non-expert SLP participants, with non-expert performance remarkable for a slightly higher Kappa when compared to expert. These findings were surprising to the investigator for several reasons.

Research has identified that training enhances rater reliability (Clain et al., 2021; Hind et al., 2009; Martin-Harris et al., 2008; Scott, Perry,& Bench, 1998), though not all training is created equal. Non-experts in the phase II final investigation were provided with several advantages over expert raters, including training with feedback of performance and on-demand definitions. Research has identified that training with feedback and definitions promotes rater reliability (Lee, Whitehill, & Ciocca, 2008). It is possible that, without the advantage of feedback, non-expert raters would have demonstrated a significantly different measure of reliability. Conversely, if expert raters were provided with on-demand definitions and feedback, it is conceivable that their measures of reliability may have outperformed non-experts.

The literature regarding rater experience and rater reliability is mixed. Rater experience has been identified as a reliability enhancing strategy (Eckberg et al., 1988; Lewis, Watterson, & Houghton, 2003; Zarkada & Regan, 2018). Phase I and II investigation findings align with the works of Martin-Harris et al. (2008), whose findings indicated that training with feedback may be more influential than rater expertise.

Demographics

The statistical analysis performed for the phase II final investigation revealed that demographic variables did not contribute to the prediction of percent accuracy. All demographic variables were not statistically significant in the regression model. These findings differed from prior literature that identified a relationship between clinical demographics and rater performance (Eckberg et al., 1988; Lewis, Watterson, & Houghton, 2003; Vose et al., 2018), Zarkada & Regan, 2018).

The investigator attributes these inconsistencies to several factors. The investigator examined the ability of a finite set of clinician demographics to predict rater performance. The demographics predictor variables presented in this investigation were not an exhaustive list of all demographics in existence. Rather, demographic characteristics with a reasonable suspicion of possible contribution to this work were included in phase II final investigation. It remains possible that clinician demographic characteristics not examined in the investigation may predict rater performance.

YSP training was provided to all study participants. Given the reported strength of training as described above, it may be reasonable to assume that training is more influential than demographic characteristics. This is consistent with prior research performed by Martin-Harris et al. (2008).

It may be interesting to also consider the research related to YSP administration and interpretation when performed by allied healthcare providers. Prior studies by Warner et al. (2014) and Schwarz et al. (2020) reported the positive operational impact and strong accuracy of swallow screening when performed by allied health professionals.

Of note, both studies emphasized training and competency evaluation as components of their research methods.

Practical Implications of the Research

Psychometrics: Maintaining Instrument Reliability Through Training

The YSP is a robust, validated swallow screening method to rapidly and accurately identify individuals at risk for swallowing impairment. A review of the literature identified that a combination of group discussions, training with consensus, feedback on performance, and clear definitions have emerged as efficacious training effects to optimize rater reliability (Clain et al., 2021; Lee, Whitehill, & Ciocca, 2008; Hind et al., 2009; Martin-Harris et al., 2008; Scott, Perry, & Bench, 1998; Stoeckli et al., 2003).

Training has been identified as an impactful contributor to the reliability of rater judgments. The current investigation has identified that training appears to be the more influential than experience or the clinical skills and certifications of SLPs. With that, training is not a component of the YSP. In light of this and in the absence of training, the investigator posits that clinicians may unintentionally deviate from the YSP guidelines resulting in degraded instrument psychometrics. With the support of the existing body of literature (Clain et al., 2021; Hind et al., 2009; Martin-Harris et al., 2008; Scott, Perry,& Bench, 1998), the investigator asserts that formalized training would likely preserve the psychometric properties of the YSP.

Competency-Based Training

Use of the Yale Swallow Protocol is not reserved exclusively for Speech-Language Pathologists. Prior studies by Warner et al. (2014) and Schwarz et al., (2020)

both reported the positive operational impact of swallow screening when performed by allied health professionals. Both studies reported participant training and competency evaluation. Competency-based training refers to the evaluative process by which the knowledge of skills are demonstrated. The competency-based education model offers an alternative to the curriculum-assessment format of the more traditional education model (Gruppen, Mangrulkar, &Kolars, 2012). Rather, "learning objectives often focus on that the learner should 'know' whereas competencies focus(es) on what the learner should be able 'to do" (Gruppen, Mangrulkar, &Kolars, 2012, p. 1). As a result, the competency-based education model evaluates the knowledge and skills of students via didactic demonstrations of educational targets with the consideration of learners' individuality and clearly communicates expectations.

While Speech-Language Pathologists have an ethical obligation to only perform services for which they are competent, the Yale Swallow Protocol does not require nor offer training prior to clinical administration. The lack of required training may contribute to clinician protocol modification (internal standards) and subsequent psychometric degradation, with potential impact on clinical findings. Pre-administration protocol training is suggested to ensure protocol adherence and preserve instrument psychometrics.

Task Complexity

The current study identified two CSVs that may negatively impact Yale Swallow Protocol interpretation. CSV four (repeated instruction) identified limited rater adherence to the protocol interpretation guidelines. Data from the current investigation identified

that many participants endorsing current WST use incorrectly rated this clinical scenario. The investigator attributed this to incomplete knowledge of the Yale Swallow Protocol.

Participant judgments of CSV five (laryngeal response) also yielded great variation. The investigator attributes the differences observed in the CSV judgments to the complexity of the clinical behavior represented on the respective CSVs and highlights the subjectivity of interpreting a laryngeal response, especially in the absence of clear definitions. The investigator asserts that clear definitions, as encouraged by Stoeckli et al. (2003), would likely assist in the consistency of rater judgments for this clinical scenario.

Evidenced-Based and Standardized Tools

The certification, training, and skills of Speech-Language Pathologists were examined in the final phase (II) of the current investigation. When examining the self-reported demographic features of participants, several interesting factors were identified.

Recall the literature reported earlier in this paper regarding the practices of SLPs, which identified the lack of standardization as a central issue (Carnaby & Harenberg, 2013; McCullough et al., 2000; Vose et al., 2018), significant variation regarding dysphagia evaluation (Martino et al., 2009), clinical decision-making (Vose et al., 2018), and rehabilitation practices (Carnaby & Harenberg, 2013). This variation in practices can be attributed, at least in part, to non-validated self-generated methods, which have been shown to possess poor diagnostic accuracy (McCullough et al., 2001). Clinicians have been encouraged to implement validated screening methods over independently established "facility-specific" protocols, considering the diagnostic weaknesses of the latter (Donovan et al., 2013).

Consistent with the existing body of literature reporting variation in practices, the current investigation identified a relatively small number of dysphagia-practicing clinicians who self-reported the implementation of validated dysphagia modalities. Surprisingly, the phase II (final) investigation identified low reported frequencies of validated dysphagia-care methods among dysphagia-practicing SLPs. Of the SLPs that endorsed the provision of dysphagia-care services, only 65% endorsed WST use. The McNeill Dysphagia Therapy Program (MDTP) is the only validated dysphagia rehabilitation protocol; only 18% of dysphagia-practicing clinicians report MDTP certification. Of the participants that reported performing Modified Barium Swallow Study skills, only 34% of participants endorse MBSImP training.

The investigator is cautious to avoid erroneous overinterpretation of the above findings and extend to the entire professional population. Within this study sample, it is noteworthy to at least mention the observed low frequency of dysphagia-specific evidence-based skills and certifications among a professional population responsible for efficacious service delivery. Considering there are a limited set of dysphagia-specific evidence-based trainings and certifications in existence, one might anticipate greater reported use of the (limited) validated dysphagia methods.

Emphasis on Aspiration

The YSP is a valid and reliable tool to identify aspiration however it has not been validated for the identification of dysphagia. While dysphagia and aspiration may be used interchangeably in the literature, they are not synonymous and can (and do) occur independently. The use of a validated swallow screening measure that only includes an investigation of aspiration may fail to identify individuals with dysphagia without

aspiration. The failure to incorporate valuable patient-reported quality of life data may withhold holistic and comprehensive swallow screening practices. To address this, clinicians may consider validated screening methods that are sensitive to both dysphagia and aspiration.

Limitations

This study is not without limitations. Statistical analyses were limited by multiple constraints that caused several mathematical limitations. First, the scale of study variables challenged the statistical analysis. The use of categorical variables limited the analyses available to the investigator. In addition, the data were abnormally distributed, which may have impacted the investigator's ability to meet the normality assumption as required by parametric statistics.

This study included a small sample size which may not be generalized to the greater population at large. Although videos were randomized, it is possible that testing effects and/or respondent fatigue impacted data.

Not all participants were presented with the same CSVs during the reliability module, which resulted in a segmented sample. The investigator recommends future studies consider a greater number of stimuli to equalize the distribution.

Finally, the investigator served as the working standard for YSP CSV ratings. This method is consistent with past precedent established in the literature (Martin-Harris et al., 2008). The current investigation planned the use of expert rater performance as the working standard (AM), however the limited adherence to the reference standard (Leder & Suiter, 2014) altered this plan. A consensus panel conducted with expert raters prior to formal data collection may optimize procedures.

Implications for Future Research

Future replication studies may wish to integrate two recommendations. First, investigators may wish to extend identical training with feedback to all participant groups. Next, a panel consensus training prior to data collection should be considered. To maximize statistical possibilities, future replication studies should consider increasing participant sample size and the quantity of stimuli presented.

Conclusion

The current investigation confirmed Speech-Language Pathologists' ability to accurately and reliability interpret 3-ounce WSCs presented across a variety of CSVs. SLP expertise did not influence rating accuracy or reliability. Although it was not part of the initial research questions, training with knowledge of performance, on-demand definitions, and unlimited video review may have impacted non-expert accuracy.

Other studies reported demographic information as a predictor of correct performance however, in the current investigation, demographic information did not predict non-expert rater accuracy. A larger and more heterogeneous participant sample is recommended if replication of the current investigation is desired.

Two CSVs (repeated instruction and laryngeal response) generated the lowest percent accuracy. The investigator suggests YSP interpretation guidelines be reviewed and updated with a recommendation to include training with feedback and expanded definitions.

References

Addington, W.R., Stephens, R.E., & Gilliland, K.A. (1999). Assessing the laryngeal cough reflex and the risk of developing pneumonia after stroke - An interhospital comparison. *Stroke (1970)*, *30*(6), 1203–1207. https://doi.org/10.1161/01.STR.30.6.1203

Al Hawat, A., Woisard, V., Perez-Begout, L., Sarrabère, E., Grand, S., & Puech, M. (2014). Validity of cervical auscultation in the screening for aspiration. *Rev Laryngol Otol Rhinol (Bord), 135*(2), 51-56.

Altman, D.G. (1999). *Practical statistics for medical research.* New York: Chapman & Hall/CRC Press.

American Speech-Language-Hearing Association ([ASHA], n.d.*) Practice Portal: Swallowing Screening*. Retrieved on 11/7/2022 [https://www.asha.org/practice-portal/clinical-topics/adult-dysphagia/swallowing-screening/]

Anderson, J. A., Pathak, S., Rosenbek, J. C., Morgan, R. O., & Daniels, S. K. (2016). Rapid Aspiration Screening for Suspected Stroke: Part 2: Initial and Sustained Nurse Accuracy and Reliability. *Arch Phys Med Rehabil, 97*(9), 1449-1455. doi: 10.1016/j.apmr.2016.03.024

Belafsky, P. C., Mouadeb, D. A., Rees, C. J., Pryor, J. C., Postma, G. N., Allen, J., & Leonard. R. J. (2008). Validity and reliability of the Eating Assessment Tool (EAT-10). *Ann Otol Rhinol Laryngol, 117*(12), 919-924. doi:10.1177/000348940811701210

Bhattacharyya, N. (2014). The prevalence of dysphagia among adults in the United States. *Otolaryngol Head Neck Surg, 151*(5), 765-769. doi:10.1177/0194599814549156

Borr, C., Hielscher-Fastabend, M., & Lücking, A. (2007). Reliability and validity of cervical auscultation. *Dysphagia, 22*(3), 225-234. doi:10.1007/s00455-007-9078- 3

Britton, D., Roeske, A., Ennis, S. K., Benditt, J. O., Quinn, C., & Graville, D. (2018). Utility of Pulse Oximetry to Detect Aspiration: An Evidence-Based Systematic Review. *Dysphagia, 33*(3), 282-292. doi:10.1007/s00455-017-9868-1

Brodsky, M. B., Suiter, D. M., Gonzalez-Fernandez, M., Michtalik, H. J., Frymark, T. B., Venediktov, R., & Schooling, T. (2016). Screening Accuracy for Aspiration Using Bedside Water Swallow Tests A Systematic Review and Meta-Analysis. *Chest, 150 (1),* 148-163. https://doi.org/10.1016/j.chest.2016.03.059

Campbell, G. B., Carter, T., Kring, D., & Martinez, C. (2016). Nursing Bedside Dysphagia Screen: Is it Valid? *J Neurosci Nurs, 48*(2), 75-79. doi:10.1097/jnn.0000000000000189

Cardoso, J. R., Pereira, L. M., Iversen, M. D., & Ramos, A. L. (2014). What is gold standard and what is ground truth? *Dental Press Journal of Orthodontics*, *19*(5), 27-30. https://doi.org/10.1590/2176-9451.19.5.027-030.ebo

Carnaby, G. D., & Harenberg, L. (2013). What is "usual care" in dysphagia rehabilitation: a survey of USA dysphagia practice patterns. *Dysphagia, 28(4),* 567-574. https://doi.org/10.1007/s00455-013-9467-8

Chan, K. M., & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. Journal of speech, language, and hearing research : JSLHR, 45(1), 111–126. https://doi.org/10.1044/1092-4388(2002/009)

Cheney, D. M., Siddiqui, M. T., Litts, J. K., Kuhn, M. A., & Belafsky, P. C. (2015). The Ability of the 10-Item Eating Assessment Tool (EAT-10) to Predict Aspiration Risk in Persons With Dysphagia. *Ann Otol Rhinol Laryngol, 124*(5), 351-354. doi:10.1177/0003489414558107

Clain, Alkhuwaiter, M., Davidson, K., & Martin-Harris, B. (2022). Structural Validity, Internal Consistency, and Rater Reliability of the Modified Barium Swallow Impairment Profile: Breaking Ground on a 52,726-Patient, Clinical Data Set. Journal of Speech, Language, and Hearing Research, 65(5), 1659–1670. https://doi.org/10.1044/2022_JSLHR-21-00554

Clave, P., Arreola, V., Romea, M., Medina, L., Palornera, E., & Serra-Prat, M. (2008). Accuracy of the volume-viscosity swallow test for clinical screening of oropharyngeal dysphagia and aspiration. *Clinical Nutrition, 27(6),* 806-815. https://doi.org/10.1016/j.clnu.2008.06.011

Clavé, P., & Shaker, R. (2015). Dysphagia: current reality and scope of the problem. *Nat Rev Gastroenterol Hepatol, 12*(5), 259-270. doi:10.1038/nrgastro.2015.49

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook C. Challenges with diagnoses: sketchy reference standards. J Man Manip Ther. 2012 Aug;20(3):111-2. doi: 10.1179/1066981712Z.00000000025. PMID: 23904748; PMCID: PMC3419566.

Curtis, J. A., & Troche, M. S. (2020). Handheld Cough Testing: A Novel Tool for Cough Assessment and Dysphagia Screening. *Dysphagia*. doi:10.1007/s00455-020-10097-z

Daniels, S. K., Ballo, L. A., Mahoney, M. C., & Foundas, A. L. (2000). Clinical predictors of dysphagia and aspiration risk: Outcome measures in acute stroke patients. *Archives of Physical Medicine and Rehabilitation, 81*(8), https://doi.org/1030-1033.10.1053/apmr.2000.6301

De Groof, I., Dejaeger, E., & Goeleven, A. (2004). Is pulse oximetrie een bruikbaar instrument om aspiratie op te sporen? [Is pulse oximetry a reliable tool for detection of aspiration?]. *Tijdschrift voor gerontologie en geriatrie*, *35*(4), 153–156.

Ding, R., & Logemann, J. A. (2008). Patient self-perceptions of swallowing difficulties as compared to expert ratings of videofluorographic studies. Folia Phoniatrica et Logopaedica, 60(3), 142–150. https://doi.org/10.1159/000120622

Donovan, N. J., Daniels, S. K., Edmiaston, J., Weinhardt, J., Summers, D., Mitchell, P. H., & Amer Heart Assoc Council, C. (2013). Dysphagia Screening: State of the Art Invitational Conference Proceeding From the State-of-the-Art Nursing Symposium, International Stroke Conference 2012. *Stroke, 44*(4), E24-E31. doi:10.1161/STR.0b013e3182877f57

Dudik, J. M., Kurosu, A., Coyle, J. L., & Sejdić, E. (2018). Dysphagia and its effects on swallowing sounds and vibrations in adults. *Biomed Eng Online, 17*(1), 69. doi:10.1186/s12938-018-0501-9

Edmiaston, J., Connor, L. T., Steger-May, K., & Ford, A. L. (2014). A simple bedside stroke dysphagia screen, validated against videofluoroscopy, detects dysphagia and aspiration with high sensitivity. *J Stroke Cerebrovasc Dis, 23*(4), 712-716. doi:10.1016/j.jstrokecerebrovasdis.2013.06.030

Ekberg, O., Nylander, G., Fork, F. T., Sjöberg, S., Birch-Iensen, M., & Hillarp, B. (1988). Interobserver variability in cineradiographic assessment of pharyngeal function during swallow. Dysphagia, 3(1), 46–48. https://doi.org/10.1007/BF02406279

Etges, C. L., Scheeren, B., Gomes, E., & Barbosa, L. e. R. (2014). Screening tools for dysphagia: a systematic review. *Codas, 26*(5), 343-349. doi:10.1590/2317-1782/20142014057

European Pharmacopoeia [Ph. Eur.] (04/2015). Reference standards. In *European Pharmacopoeia* (9.0, pp. 733-736). European Pharmacopoeia.

Exley, C. (2000). Pulse oximetry as a screening tool in detecting aspiration. *Age Ageing, 29*(6), 475-476. doi:10.1093/ageing/29.6.475

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160.

Festic, Soto, J. S., Pitre, L. A., Leveton, M., Ramsey, D. M., Freeman, W. D., Heckman, M. G., & Lee, A. S. (2016). Novel Bedside Phonetic Evaluation to Identify Dysphagia and Aspiration Risk. Chest., 149(3), 649–659. https://doi.org/10.1378/chest.15-0789

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.

Frowen, J. J., Cotton, S. M., & Perry, A. R. (2008). The stability, reliability, and validity of videofluoroscopy measures for patients with head and neck cancer. *Dysphagia*, *23*(4), 348–363. https://doi.org/10.1007/s00455-008-9148-1

Garand, K. L. F., McCullough, G., Crary, M., Arvedson, J. C., & Dodrill, P. (2020). Assessment Across the Life Span: The Clinical Swallow Evaluation. *Am J Speech Lang Pathol, 29*(2s), 919-933. doi:10.1044/2020_ajslp-19-00063

Goldstone, R. L. (1998). Perceptual learning. Annual Review of Psychology, 49, 585–612.

Groves-Wright, K. J., Boyce, S., & Kelchner, L. (2010). Perception of wet vocal quality in identifying penetration/aspiration during swallowing. *J Speech Lang Hear Res, 53*(3), 620-632. doi:10.1044/1092-4388(2009/08-0246)

Gruppen, L. D., Mangrulkar, R. S., & Kolars, J. C. (2012). The promise of competency-based education in the health professions for improving global health. Human resources for health, 10, 43. https://doi.org/10.1186/1478-4491-10-43

Guillén-Solà, A., Marco, E., Martínez-Orfila, J., Donaire Mejías, M. F., Depolo Passalacqua, M., Duarte, E., & Escalada, F. (2013). Usefulness of the volume-viscosity swallow test for screening dysphagia in subacute stroke patients in rehabilitation income. *NeuroRehabilitation, 33*(4), 631-638. doi:10.3233/nre-130997

Hara, K., Tohara, H., Wada, S., Iida, T., Ueda, K., & Ansai, T. (2014). Jaw-opening force test to screen for Dysphagia: preliminary results. *Arch Phys Med Rehabil, 95*(5), 867-874. doi:10.1016/j.apmr.2013.09.005

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. Journal of Applied Behavior Analysis, 10, 103–116.

Hinchey, J. A., Shephard, T., Furie, K., Smith, D., Wang, D., Tonn, S., & Stroke Practice Improvement, N. (2005). Formal dysphagia screening Protocols prevent pneumonia. *Stroke, 36*(9), 1972-1976. doi:10.1161/01.STR.0000177529.86868.8d

Hind, Gensler, G., Brandt, D. K., Miller Gardner, P. J., Blumenthal, L., Gramigna, G. D., Kosek, S., Lundy, D., McGarvey-Toler, S., Rockafellow, S., Sullivan, P. A., Villa, M., Gill, G. D., Lindblad, A. S., Logemann, J. A., & Robbins, J. (2009). Comparison of Trained Clinician Ratings with Expert Ratings of Aspiration on Videofluoroscopic Images from a Randomized Clinical Trial. Dysphagia, 24(2), 211–217. https://doi.org/10.1007/s00455-008-9196-6

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.

Hughes, T.A., & Wiles, C.M. (1996) Clinical measurement of swallowing in health and in neurogenic dysphagia. *QJM: An International Journal of Medicine, 89* (2), 109–116, https://doi.org/10.1093/qjmed/89.2.109

IBM Corp. (2021) SPSS Statistics for Windows (Version 28.0.1). Armonk, N.Y.

Jacobsen, K.H. (2017). *Introduction to health research: a practical guide (*2nd ed.). Jones & Bartlett learning

Kidd, D., Lawson, J., Nesbitt, R., & MacMahon, J. (1993). Aspiration in acute stroke: A clinical study with videofluoroscopy. *Quarterly Journal of Medicine, 86*(12), 825–829.

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159-174.

Langmore, S. E., Terpenning, M. S., Schork, A., Chen, Y., Murray, J. T., Lopatin, D., & Loesche, W. J. (1998). Predictors of aspiration pneumonia: how important is dysphagia? *Dysphagia, 13*(2), 69-81. doi:10.1007/pl00009559

Leder, S. B. Use of arterial oxygen saturation, heart rate, and blood pressure as indirect objective physiologic markers to predict aspiration. Dysphagia. 2000 Fall;15(4):201-5. doi: 10.1007/s004550000028. PMID: 11014882.

Leder, S. B., & Suiter, D. M. (2014). *The Yale Swallow Protocol: An evidence-based approach to decision making*. Springer. Springer International Publishing. Switzerland 2014

Leder, S. B., Suiter, D. M., & Green, B. G. (2011). Silent aspiration risk is volume-dependent. *Dysphagia, 26*(3), 304-309. doi:10.1007/s00455-010-9312-2

Leder, S. B., Suiter, D. M., Warner, H. L., Acton, L. M., & Siegel, M. D. (2012). Safe initiation of oral diets in hospitalized patients based on passing a 3-ounce (90 cc) water swallow challenge protocol. *Qjm, 105*(3), 257-263. doi:10.1093/qjmed/hcr193

Leder, S. B., Suiter, D. M., Warner, H. L., Acton, L. M., & Swainson, B. A. (2012). Success of recommending oral diets in acute stroke patients based on passing a 90-cc water swallow challenge protocol. *Top Stroke Rehabil, 19*(1), 40-44. doi:10.1310/tsr1901-40

Lee, A., Whitehill, T. L., & Ciocca, V. (2009). Effect of listener training on perceptual judgement of hypernasality. Clinical Linguistics & Phonetics, 23(5), 319–334. https://doi.org/10.1080/02699200802688596

Lee, J. H., & Choi, S. Y. (2020). Criteria to assess tongue strength for predicting penetration and aspiration in patients with stroke having dysphagia. *Eur J Phys Rehabil Med, 56*(4), 375-385. doi:10.23736/s1973-9087.20.06180-8

Lee, J. Y., Kim, D. K., Seo, K. M., & Kang, S. H. (2014). Usefulness of the simplified cough test in evaluating cough reflex sensitivity as a screening test for silent aspiration. *Ann Rehabil Med, 38*(4), 476-484. doi:10.5535/arm.2014.38.4.476

Lewis, K. E., Watterson, T. L., & Houghton, S. M. (2003). The influence of listener experience and academic training on ratings of nasality. *Journal of communication disorders*, *36*(1), 49–58. https://doi.org/10.1016/s0021-9924(02)00134-x

Lim, J. Y., Yoo, Y. H., Park, C. H., Joa, K. L., & Jung, H. Y. (2020). Use of the maximal phonation test for the screening of dysphagia in stroke patients: a preliminary study. *Eur J Phys Rehabil Med, 56*(1), 41-46. doi:10.23736/s1973-9087.19.05818-0

Logemann, J. A., Veis, S., & Colangelo, L. (1999). A screening procedure for oropharyngeal dysphagia. *Dysphagia, 14*(1), 44–51. https://doi.org/10.1007/PL00009583

Lowell, S. Y., Kelley, R. T., Busekroos, L., Voleti, R., Hosbach-Cannon, C. J., Colton, R. H., & Mihaila, D. (2017). The effect of anchors on reliability of endoscopic

tremor ratings. The Laryngoscope, 127(2), 411–416. https://doi.org/10.1002/lary.26034

Malandraki, G. A., Hind, J. A., Gangnon, R., Logemann, J. A., & Robbins, J. (2011). The utility of pitch elevation in the evaluation of oropharyngeal Dysphagia: preliminary findings. *Am J Speech Lang Pathol, 20*(4), 262-268. doi:10.1044/1058-0360(2011/10-0097)

Mari, F., Matei, M., Ceravolo, M. G., Pisani, A., Montesi, A., & Provinciali, L. (1997). Predictive value of clinical indices in detecting aspiration in patients with neurological disorders. *Journal of Neurology, Neurosurgery, and Psychiatry, 63*(4), 456–460.

Marian, T., Schröder, J., Muhle, P., Claus, I., Oelenberg, S., Hamacher, C., . . . Dziewas, R. (2017). Measurement of Oxygen Desaturation Is Not Useful for the Detection of Aspiration in Dysphagic Stroke Patients. *Cerebrovasc Dis Extra, 7*(1), 44-50. doi:10.1159/000453083

Martin-Harris, B. (2015). *Standardized Training in Swallowing Physiology – Evidence-Based Assessment Using the Modified Barium Swallow Impairment Profile (MBSImP) Approach*. Northern Speech Services, Inc.

Martin-Harris, B., Brodsky, M. B., Michel, Y., Castell, D. O., Schleicher, M., Sandidge, J., Maxwell, R., & Blair, J. (2008). MBS measurement tool for swallow impairment--MBSImp: establishing a standard. *Dysphagia*, *23*(4), 392–405. https://doi.org/10.1007/s00455-008-9185-9

Martino, R., Silver, F., Teasell, R., Bayley, M., Nicholson, G., Streiner, D. L., & Diamant, N. E. (2009). The Toronto Bedside Swallowing Screening Test (TOR-

BSST) Development and Validation of a Dysphagia Screening Tool for Patients With Stroke. *Stroke, 40 (2*), 555-561. https://doi.org/10.1161/strokeaha.107.510370

Matsuo, K., & Palmer, J. B. (2008). Anatomy and Physiology of Feeding and Swallowing –Normal and Abnormal. Physical medicine and rehabilitation clinics of North America, 19(4), 691. https://doi.org/10.1016/j.pmr.2008.06.001

McCullough, G. H., Wertz, R. T., Rosenbek, J. C., Mills, R. H., Ross, K. B., & Ashford, J. R. (2000). Inter- and intrajudge reliability of a clinical examination of swallowing in adults. *Dysphagia, 15*(2), 58-67. doi:10.1007/s004550010002

McCullough, G. H., Wertz, R. T., & Rosenbek, J. C. (2001). Sensitivity and specificity of clinical/bedside examination signs for detecting aspiration in adults subsequent to stroke. *Journal of Communication Disorders, 34(1-2),* 55-72. https://doi.org/10.1016/s0021-9924(00)00041-1

McHorney, C. A., Robbins, J., Lomax, K., Rosenbek, J. C., Chignell, K., Kramer, A. E., & Bricker, D. E. (2002). The SWAL-QOL and SWAL-CARE outcomes tool for oropharyngeal dysphagia in adults: III. Documentation of reliability and validity. *Dysphagia, 17*(2), 97-114. doi:10.1007/s00455-001-0109-1

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Morrissey, A.M. [AM]. (n.d.). *AM* [YouTube channel] YouTube. Retrieved February 14, 2022, from https://www.youtube.com/AM.

Nathadwarawala, K.M., Nicklin, J., & Wiles, C. M. (1992). A timed test of swallowing capacity for neurological patients. Journal of Neurology, Neurosurgery and Psychiatry, 55(9), 822–825. https://doi.org/10.1136/jnnp.55.9.822

Nielsen, A. H., Gow, N. D., & Svenningsen, H. (2021). Translation and adaption of the Yale Swallow Protocol for a Danish intensive care setting. *Scandinavian journal of caring sciences*, *35*(4), 1290–1300. https://doi.org/10.1111/scs.12950

Perry, L. (2001). Screening swallowing function of patients with acute stroke. Part one: Identification, implementation and initial evaluation of a screening tool for use by nurses. *J Clin Nurs, 10*(4), 463-473. doi:10.1046/j.1365-2702.2001.00501.x

Qualtrics software, Version September 2021 of Qualtrics. Copyright 2021. Provo, Utah, USA. Available at https://www.qualtrics.com

Rajappa, A. T., Soriano, K. R., Ziemer, C., Troche, M. S., Malandraki, J. B., & Malandraki, G. A. (2017). Reduced Maximum Pitch Elevation Predicts Silent Aspiration of Small Liquid Volumes in Stroke Patients. *Front Neurol, 8*, 436. doi:10.3389/fneur.2017.00436

Rameau A, Katz P, Andreadis K, et al. Clarifying Inaccurate Terminology: The Important Difference Between Dysphagia and Swallowing Dysfunction. Foregut. 2022;2(1):11-17. doi:10.1177/26345161211072761

Roaché, D. (2017). Intercoder reliability techniques: percent agreement. In M. Allen (Ed.), The sage encyclopedia of communication research methods (pp. 752-752). SAGE Publications, Inc, https://dx.doi.org/10.4135/9781483381411.n260

Roberts, M. Y., Sone, B. J., Zanzinger, K. E., Bloem, M. E., Kulba, K., Schaff, A., . . . Goldstein, H. (2020). Trends in Clinical Practice Research in ASHA Journals:

2008-2018. *Am J Speech Lang Pathol, 29*(3), 1629-1639. doi:10.1044/2020_ajslp-19-00011

Rofes, L., Arreola, V., & Clavé, P. (2012). The volume-viscosity swallow test for clinical screening of dysphagia and aspiration. *Nestle Nutr Inst Workshop Ser, 72*, 33-42. doi:10.1159/000339979

Rosenbek, J. C., Robbins, J. A., Roecker, E. B., Coyle, J. L., & Wood, J. L. (1996). A penetration-aspiration scale. *Dysphagia*, *11*(2), 93–98. https://doi.org/10.1007/BF00417897

Runions, S., Rodrigue, N., & White, C. (2004). Practice on an acute stroke unit after implementation of a decision-making algorithm for dietary management of dysphagia. *Journal of Neuroscience Nursing, 36*(4), 200–207. https://doi.org/10.1097/01376517-200408000-00006

Ryu, J. S., Park, S. R., & Choi, K. H. (2004). Prediction of laryngeal aspiration using voice analysis. *American Journal of Physical Medicine and Rehabilitation, 83*(10), 753–757. https://doi.org/10.1097/01.phm.0000140798.97706.a5

Sarraf Shirazi, S., & Moussavi, Z. (2012). Silent aspiration detection by breath and swallowing sound analysis. *Annu Int Conf IEEE Eng Med Biol Soc, 2012*, 2599-2602. doi:10.1109/embc.2012.6346496

Sato, M., Tohara, H., Iida, T., Wada, S., Inoue, M., & Ueda, K. (2012). Simplified cough test for screening silent aspiration. *Arch Phys Med Rehabil, 93*(11), 1982-1986. doi:10.1016/j.apmr.2012.05.016

Schwarz, M., Ward, E. C., Cornwell, P., Coccetti, A., D'Netto, P., Smith, A., & Morley-Davies, K. (2020). Exploring the Validity and Operational Impact of Using Allied

Health Assistants to Conduct Dysphagia Screening for Low-Risk Patients Within the Acute Hospital Setting. *American journal of speech-language pathology*, *29*(4), 1944–1955. https://doi.org/10.1044/2020_AJSLP-19-00060

Scott A, Perry A, Bench J: A study of interrater reliability when using videofluoroscopy as an assessment of swallowing. Dysphagia 13:223–227, 1998

Shadish, W.R., Cook, T. D. & Campbell, D.T. (2002) *Experimental and Quasi-Experimental Design* . Wadsworth Cengage.

Silbergleit, A. K., Schultz, L., Jacobson, B. H., Beardsley, T., & Johnson, A. F. (2012). The Dysphagia handicap index: development and validation. *Dysphagia, 27*(1), 46-52. doi:10.1007/s00455-011-9336-2

Šimundić, M. (2009). Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC*, *19*(4), 203-211. https://doi.org/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975285/

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*, 799-811.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4). http://pareonline.net/getvn.asp?v=9&n=4

Stoeckli, Huisman, T. A. G. M., Seifert, B., & Martin-Harris, B. J. W. (2003). Interrater reliability of videofluoroscopic swallow evaluation. Dysphagia, 18(1), 53–57. https://doi.org/10.1007/s00455-002-0085-0

Stoltzfus J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Stroud, A. E., Lawrie, B. W., & Wiles, C. M. (2002). Inter- and intra-rater reliability of cervical auscultation to detect aspiration in patients with dysphagia. *Clin Rehabil, 16*(6), 640-645. doi:10.1191/0269215502cr533oa

Suiter, D. (2018). Dysphagia Screening: Challenges and Controversies. *Perspectives of the ASHA Special Interest Groups, 3*(13), 82-88. doi:10.1044/persp3.SIG13.82

Suiter, D. M., Daniels, S. K., Barkmeier-Kraemer, J. M., & Silverman, A. H. (2020). Swallowing Screening: Purposefully Different From an Assessment Sensitivity and Specificity Related to Clinical Yield, Interprofessional Roles, and Patient Selection. *American Journal of Speech-Language Pathology, 29*(2), 979-991. doi:10.1044/2020_ajslp-19-00140

Suiter, D. M., & Leder, S. B. (2008). Clinical utility of the 3-ounce water swallow test. *Dysphagia, 23*(3), 244-250. doi:10.1007/s00455-007-9127-y

Suiter, D. M., Sloggy, J., & Leder, S. B. (2014). Validation of the Yale Swallow Protocol: a prospective double-blinded videofluoroscopic study. *Dysphagia, 29*(2), 199-203. doi:10.1007/s00455-013-9488-3

Vose, A. K., Kesneck, S., Sunday, K., Plowman, E., & Humbert, I. (2018). A Survey of Clinician Decision Making When Identifying Swallowing Impairments and Determining treatment. *J Speech Lang Hear Res, 61*(11), 2735-2756. doi:10.1044/2018_jslhr-s-17-0212

Wakasugi, Y., Tohara, H., Nakane, A., Murata, S., Mikushi, S., Susa, C., . . . Uematsu, H. (2014). Usefulness of a handheld nebulizer in cough test to screen for silent aspiration. *Odontology, 102*(1), 76-80. doi:10.1007/s10266-012-0085-y

Wallace, A. L., Middleton, S., & Cook, I. J. (2000). Development and validation of a self-report symptom inventory to assess the severity of oral-pharyngeal dysphagia. *Gastroenterology, 118*(4), 678-687. doi:10.1016/s0016-5085(00)70137-5

Wang, T. G., Chang, Y. C., Chen, S. Y., & Hsiao, T. Y. (2005). Pulse oximetry does not reliably detect aspiration on videofluoroscopic swallowing study. *Arch Phys Med Rehabil, 86*(4), 730-734. doi:10.1016/j.apmr.2004.10.021

Wangen, T., Hatlevig, J., Pifer, G., & Vitale, K. (2019). Preventing Aspiration Complications: Implementing a Swallow Screening Tool. *Clin Nurse Spec, 33*(5), 237-243. doi:10.1097/nur.0000000000000471

Ward, M., Skelley-Ashford, M., Brown, K., Ashford, J., & Suiter, D. (2020). Validation of the Yale Swallow Protocol in Post-Acute Care: A Prospective, Double-Blind, Multirater Study. *Am J Speech Lang Pathol*, 1-7. doi:10.1044/2020_ajslp-19-00147

Warms, T., & Richards, J. (2000). "Wet Voice" as a predictor of penetration and aspiration in oropharyngeal dysphagia. *Dysphagia, 15*(2), 84-88. doi:10.1007/s004550010005

Warner, H. L., Suiter, D. M., Nystrom, K. V., Poskus, K., & Leder, S. B. (2014). Comparing accuracy of the Yale swallow protocol when administered by

registered nurses and speech-language pathologists. *Journal of clinical nursing*, *23*(13-14), 1908–1915. https://doi.org/10.1111/jocn.12340

Weinhardt, J., Hazelett, S., Barrett, D., Lada, R., Enos, T., & Keleman, R. (2008). Accuracy of a Bedside Dysphagia Screening: A Comparison of Registered Nurses and Speech Therapists. *Rehabilitation Nursing, 33*(6), 247-252. doi:10.1002/j.2048-7940.2008.tb00236.x

Zaidi, N. H., Smith, H. A., King, S. C., Park, C., O'Neill, P. A., & Connolly, M. J. (1995). Oxygen desaturation on swallowing as a potential marker of aspiration in acute stroke. *Age Ageing, 24*(4), 267-270. doi:10.1093/ageing/24.4.267

Zarkada, A., & Regan, J. (2018). Inter-rater Reliability of the Dysphagia Outcome and Severity Scale (DOSS): Effects of Clinical Experience, Audio-Recording and Training.

Dysphagia, *33*(3), 329–336. https://doi.org/10.1007/s00455-017-9857-4

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High Agreement and High Prevalence: The Paradox of Cohen's Kappa. *The open nursing journal*, *11*, 211–218. https://doi.org/10.2174/1874434601711010211

Appendix A – Yale Swallow Protocol

Chapter 13
Yale Swallow Protocol Administration Forms

Administration Form 1.

Yale Swallow Protocol

Step 1: Exclusion Criteria

__ **Protocol Deferred: NO risk factors for aspiration.**

Protocol deferred if any YES answer to the following criteria

Yes	No	
__	__	Unable to remain alert for testing
__	__	No thin liquids due to preexisting dysphagia
__	__	Head-of-Bed restricted to <30°
__	__	Tracheotomy tube present
__	__	Nil-per-os order for medical/surgical reason

If a patient's clinical status changes resulting in a new risk for aspiration re-administer protocol before oral intake of food or medicine.

S.B. Leder and D.M. Suiter, *The Yale Swallow Protocol: An Evidence-Based Approach to Decision Making*, DOI 10.1007/978-3-319-05113-0_13,

150 13. Yale Swallow Protocol Administration Forms

Administration Form 2.*

Yale Swallow Protocol

Step 2: Administration Instructions

Perform protocol if patient is an aspiration risk and ALL Step 1 boxes are checked NO

- **Brief Cognitive Screen**[a]: __ What is your name? __ Open your mouth
  __ Where are you right now? __ Stick out your tongue
  __ What year is it? __ Smile
- **Oral-Mechanism Examination**[b]: __ Labial closure
  __ Lingual range of motion
  __ Facial symmetry (smile/pucker)
- **3-Ounce Water Swallow Challenge**[c]:
- Sit patient upright at 80-90° (or as high as tolerated >30°)
- Ask patient to drink the entire 3 ounces (90cc) of water from a cup or with a straw, in sequential swallows, and slow and steady but without stopping
  (Note: Cup or straw can be held by staff or patient)
- Assess patient for coughing or choking during or immediately after completion of drinking

[a,b] Information from the brief cognitive screen and oral mechanism examination provide information only on odds of aspiration risk with the 3-ounce water swallow challenge and should not be used as exclusionary criteria for screening.

[c] It is permissible to repeat the 3-ounce water swallow challenge if it is thought the patient may pass with a second attempt.

Administration Form 3.*

Yale Swallow Protocol

Step 3: Pass/Fail Criteria

Results and Recommendations

__ **PASS: Successful uninterrupted drinking of all 3 ounces of water without overt signs of aspiration (coughing/choking) either during or immediately after completion.**

- If patient passes, collaborate with MD/PA/LIP to order appropriate oral diet.
- If adequate dentition order a soft solid consistency or regular consistency diet.
- If inadequate dentition or edentulous order a liquid and puree diet.
- Consult with speech-language pathologist for other diet modifications.

__ **FAIL: Inability to drink the entire 3 ounces in sequential swallows due to interrupted drinking (stopping/starting) or patient exhibits overt signs of aspiration (coughing/choking) either during or immediately after completion.**

- If patient fails, keep nil per os (including medications) and request the MD/PA/LIP to order a consult for an instrumental swallowing evaluation by speech-language pathology.

  OR

- Continue nil-per-os status and re-administer the protocol in 24 hours if patient shows clinical improvement.
- If patient fails again request the MD/PA/LIP to order a consult for an instrumental swallowing evaluation by speech-language pathology.

Appendix B -Supplemental Materials

Pilot Study

Table _
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Percentage of Accuracy	8	64.7	88.0	77.125	6.5689
Valid N (listwise)	8				

Statistics

Percentage of Accuracy		
N	Valid	8
	Missing	1
Mean		77.125
Median		76.400
Mode		76.4
Std. Deviation		6.5689
Minimum		64.7
Maximum		88.0

Table _
Percentage of Accuracy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	64.7	1	11.1	12.5	12.5
	76.4	5	55.6	62.5	75.0
	82.3	1	11.1	12.5	87.5
	88.0	1	11.1	12.5	100.0
	Total	8	88.9	100.0	
Missing	System	1	11.1		
Total		9	100.0		

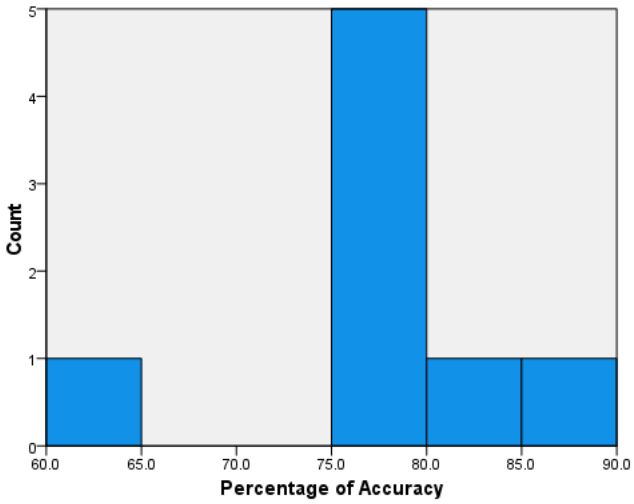


Table _

Rater accuracy

Rater	Percentage of Accuracy
Rater 1	82.3%
Rater 2	76.4%
Rater 3	76.4%
Rater 4	76.4%
Rater 5	88%
Rater 6	76.4%
Rater 7	76.4%
Rater 8	64.7%

Note. Accuracy comparing reference standard (investigator) with individual raters.

Table _: T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Percentage of Accuracy	8	77.125	6.5689	2.3225

One-Sample Test

	Test Value = 75						
			Significance			95% Confidence Interval of the Difference	
	t	df	One-Sided p	Two-Sided p	Mean Difference	Lower	Upper
Percentage of Accuracy	.915	7	.195	.391	2.1250	-3.367	7.617

Table _

Clinical Scenario Videos with Accuracy

Cup delivery method		Accuracy		Straw delivery method	Accuracy
Video number			Video number		
1. Pass	Uninterrupted drinking	100%	10. Pass	Uninterrupted drinking	**50%**
2. Fail	Delayed cough	87.5%			
3. Pass	Uninterrupted drinking	100%			
4. Pass	Repeat instruction	**50%**	11. Pass	Repeat instruction	**0%**
5. Pass	Throat clear	**0%**	12. Pass	Throat clear	**0%**
6. Fail	Interrupted drinking & coughing	100%	13. Fail	Interrupted drinking, & coughing	100%
7. Fail	Interrupted drinking & throat clearing	100%	14. Fail	Interrupted drinking & throat clearing	100%
8. Fail	Interrupted drinking, no cough nor throat clear	87.5%	15. Fail	Interrupted drinking, no cough nor throat clear	100%
9. Fail	Uninterrupted drinking with immediate cough	100%	16. Fail	Uninterrupted with immediate cough	100%
			17. Fail	Uninterrupted with immediate cough	100%

Table _

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	cupaccuracy	80.1750	8	7.48785	2.64735
	strawaccuracy	71.8750	8	5.78638	2.04579

Paired Samples Correlations

				Significance	
				One-Sided	Two-Sided
		N	Correlation	p	p
Pair 1	cupaccuracy & strawaccuracy	8	.237	.286	.572

Paired Samples Test

		Paired Differences							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Pair 1	cupaccuracy - strawaccuracy	**8.30000**	8.30748	2.93714	1.35477	15.24523	**2.826**	7	.013	**.026**

Paired Samples Effect Sizes

					95% Confidence Interval	
			Standardizer[a]	Point Estimate	Lower	Upper
Pair 1	cupaccuracy - strawaccuracy	Cohen's d	8.30748	.999	.115	1.839
		Hedges' correction	9.35314	.887	.102	1.634

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

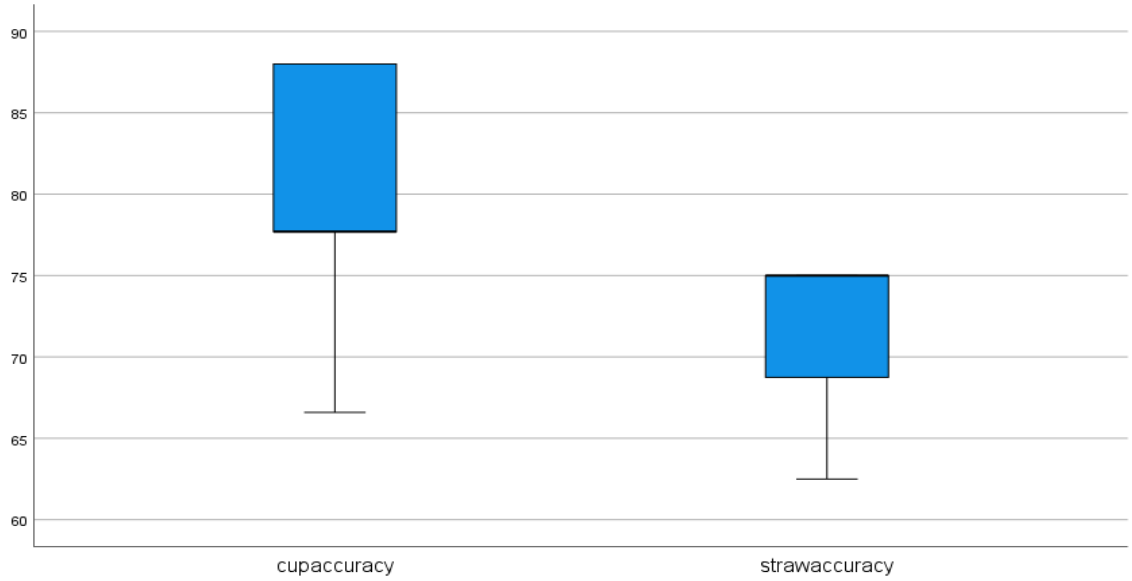


Table _

Intrarater reliability summary

Rater	Overall proportion of agreement	Strength of agreement	Cohen's Kappa Coefficient
Rater 1	100%	Perfect	κ = 1.0, 95% CI [0.0, 0.0], $p < .001$.
Rater 2	94.1%	Very good	κ = .877, 95% CI [.664, 1.09], $p < .001$.
Rater 3	100%	Perfect	κ = 1.0, 95% CI [0.0, 0.0], $p < .001$.
Rater 4	94.1%	Very good	κ = .821, 95% CI [-0.164. .560], $p < .001$.
Rater 5	100%	Perfect	κ = 1.0, 95% CI [0.0, 0.0], $p < .001$.
Rater 6	n/a	n/a	n/a
Rater 7	100%	Perfect	κ = 1.0, 95% CI [0.0, 0.0], $p < .001$.
Rater 8	94.1%	Very good	κ = .821, 95% CI [.590, 1.126], $p < .001$.

Note. Table includes Case Processing Summary, First Rating * Reliability Rating Crosstabulation, and Symmetric Measures. To adjust for the limited CI as reported by asymptomatic standard error, a 95% CI manually calculated (asymptomatic standard error x 1.96).

Table _

Overall Agreement[a]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.649	.040	16.179	.000	.570	.727

a. Sample data contains 17 effective subjects and 8 raters.

Table _

Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	.694	.615	.046	13.407	.000	.525	.704
1	.938	.743	.046	16.206	.000	.653	.833
2	.343	.318	.046	6.933	<.001	.228	.408

a. Sample data contains 17 effective subjects and 8 raters.

Table _

Cup Overall Agreement[a]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	**.721**	.063	11.453	**.000**	**.598**	**.845**

a. Sample data contains 9 effective subjects and 8 raters.

Table _

Cup Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	.803	**.721**	.063	11.453	.000	**.598**	**.845**
1	.919	**.721**	.063	11.453	.000	**.598**	**.845**

a. Sample data contains 9 effective subjects and 8 raters.

Table _

Straw Overall Agreement[a]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	**.514**	.052	9.959	**.000**	**.413**	**.616**

a. Sample data contains 8 effective subjects and 8 raters.

Table _

Straw Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	.367	**.290**	.067	4.335	<.001	.159	.421
1	.956	**.766**	.067	11.458	.000	.635	.897
2	.343	**.287**	.067	4.298	<.001	.156	.418

a. Sample data contains 8 effective subjects and 8 raters.

Final Study

Accuracy

One-Sample Statistics: One-Sample t Test

	N	Mean	Std. Deviation	Std. Error Mean	t	df	P One-Sided p	95% CI Lower	95% CI Upper	Mean Difference
Percent accuracy	150	90.0651	8.45492	.69034	13.899	149	<.001	8.2309	10.9592	9.59507

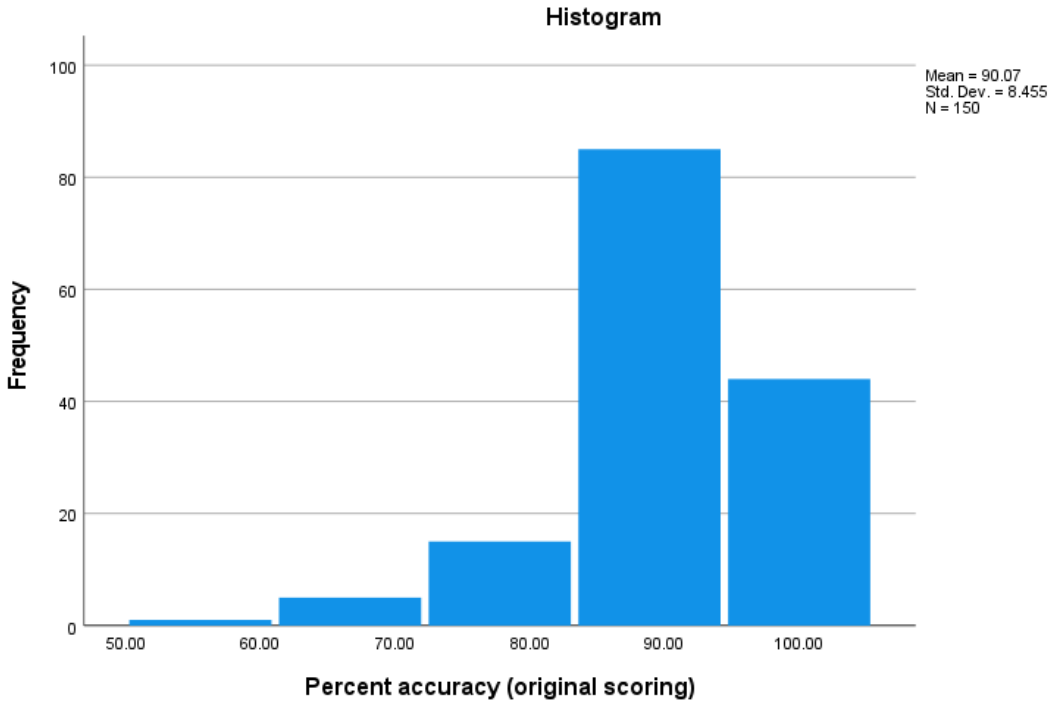
Test Value = 80.47

Table _

Descriptives

		Statistic	Std. Error
Percent accuracy	Mean	90.0651	.69034

	95% Confidence Interval for Mean	Lower Bound	88.7009	
		Upper Bound	91.4292	
	5% Trimmed Mean		90.7728	
	Median		88.8800	
	Variance		71.486	
	Std. Deviation		8.45492	
	Minimum		55.55	
	Maximum		99.99	
	Range		44.44	
	Interquartile Range		11.11	
	Skewness		-1.015	.198
	Kurtosis		1.983	.394



One-Sample Effect Sizes

			Point	95% CI	
		Standardizer[a]	Estimate	Lower	Upper
Percent accuracy	Cohen's d	8.45492	1.135	.928	1.339

	Hedges' correction	8.49778	1.129	.924	1.332

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

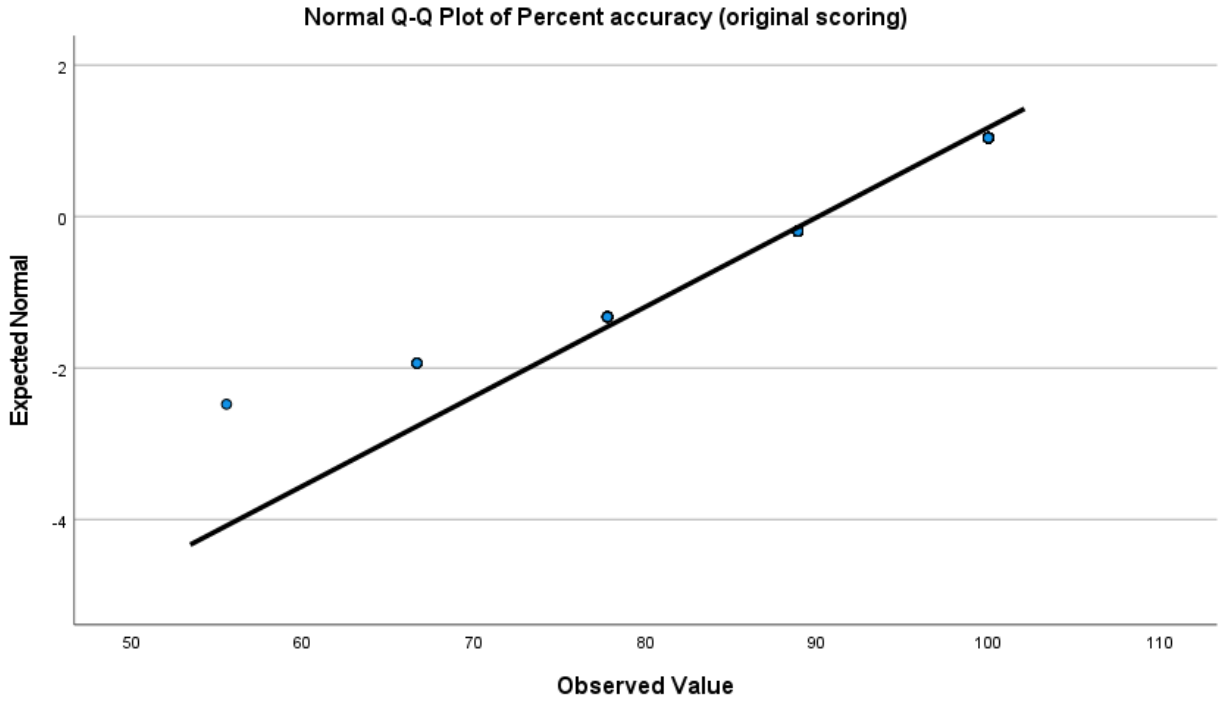
Case Processing Summary

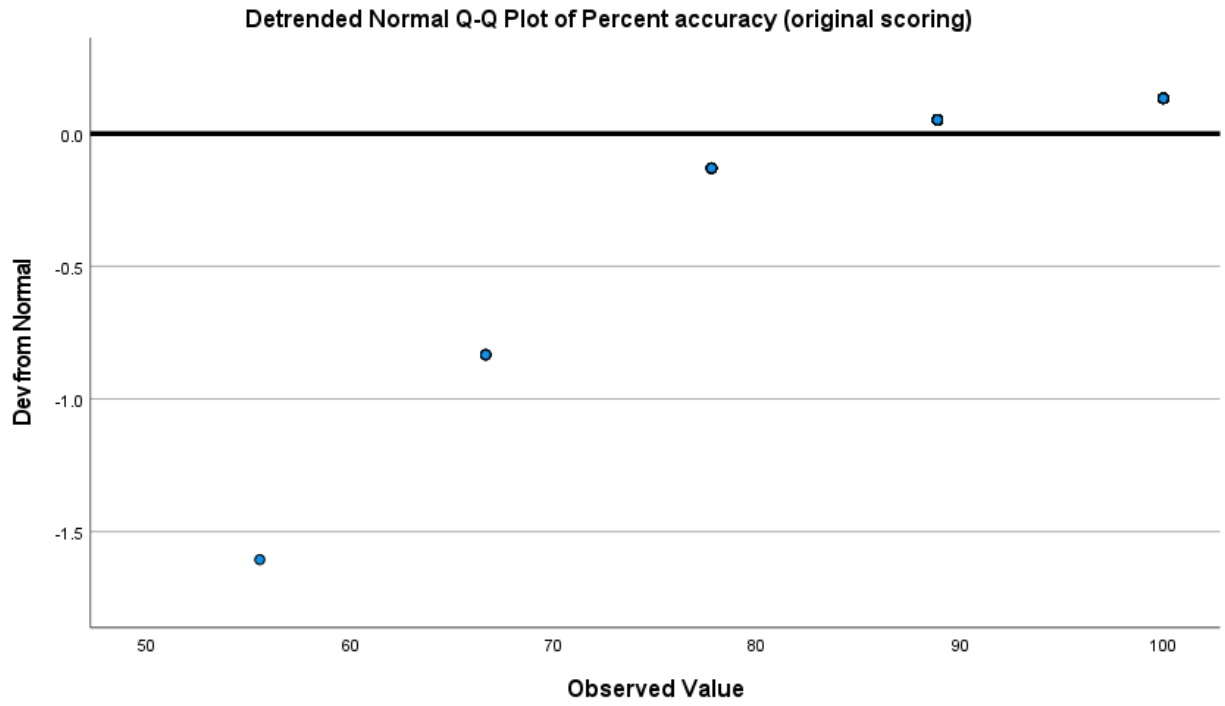
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Percent accuracy	150	100.0%	0	0.0%	150	100.0%

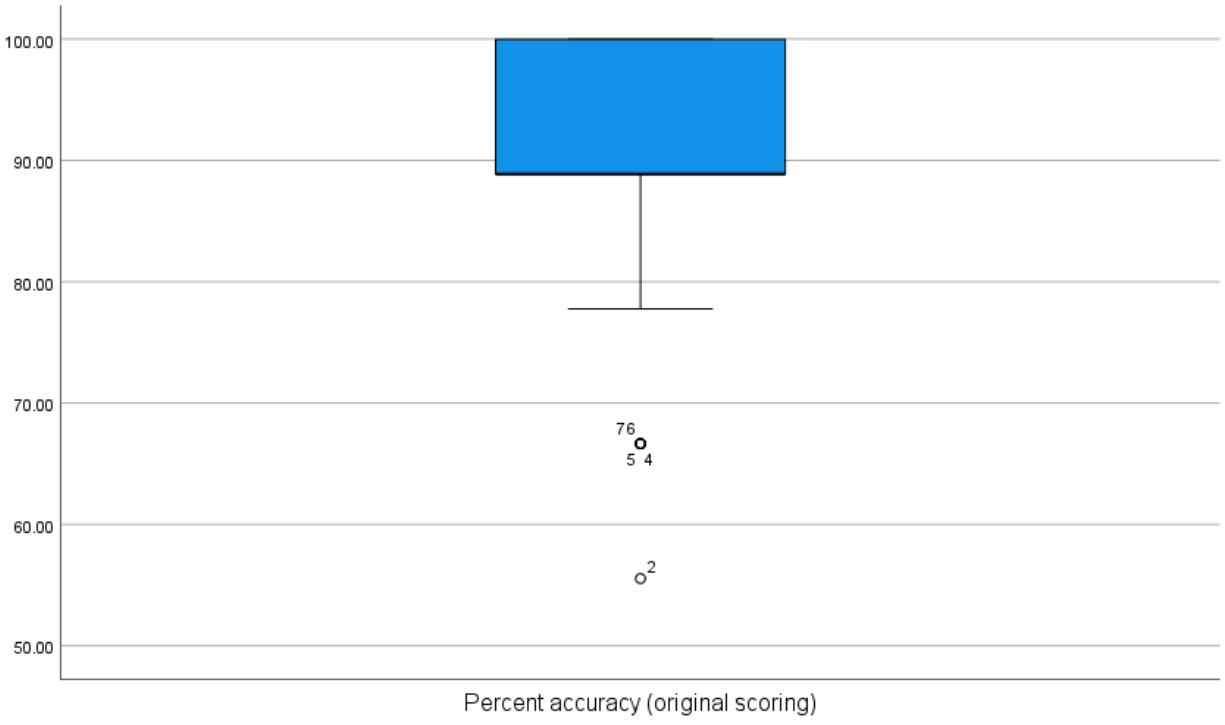
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Percent accuracy	.304	150	<.001	.785	150	<.001

a. Lilliefors Significance Correction







Hypothesis Test Summary

Null Hypothesis	Test	Sig.[a,b]	Decision

1	The median of Percent accuracy equals 77.70.	One-Sample Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

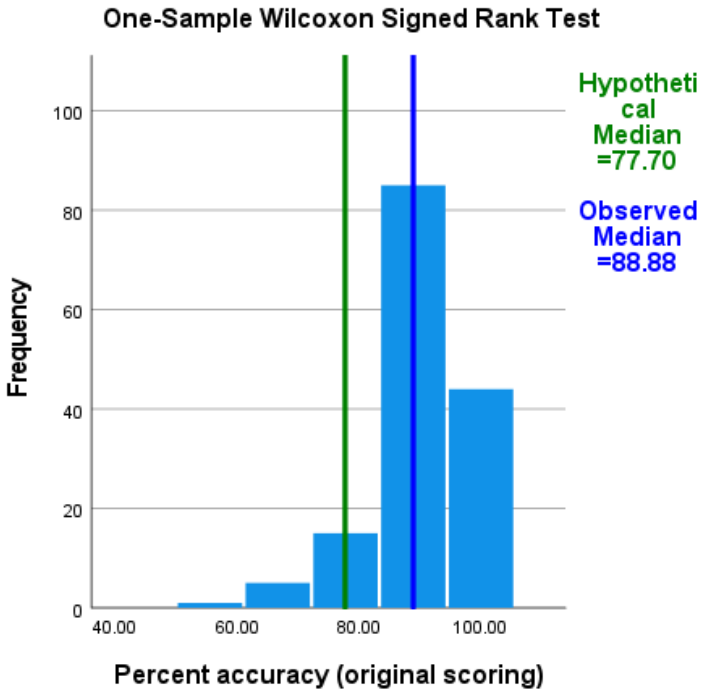
a. The significance level is .050.

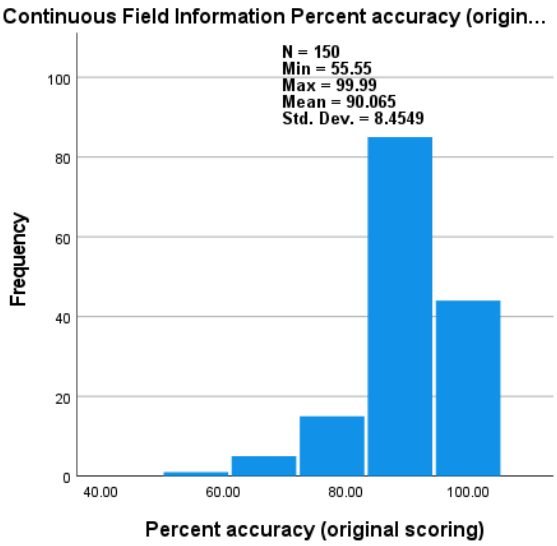
b. Asymptotic significance is displayed.

Table _

One-Sample Wilcoxon Signed Rank Test Summary

Total N	150
Test Statistic	11129.000
Standard Error	519.066
Standardized Test Statistic	10.531
Asymptotic Sig.(2-sided test)	.000






```
FREQUENCIES VARIABLES=Video_1 Video_2 Video_3 Video_4
Video_5 Video_6 Video_7 Video_8 Video_9
  /ORDER=ANALYSIS.
```

Frequencies

Notes

Output Created		08-NOV-2022 13:08:18
Comments		
Input	Data	\\files\users\amorrissey\Desktop\Final study material\SPSS 6-15-22.sav
	Active Dataset	DataSet4
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	150
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data.

Syntax		FREQUENCIES VARIABLES=Video_1 Video_2 Video_3 Video_4 Video_5 Video_6 Video_7 Video_8 Video_9 /ORDER=ANALYSIS.
Resources	Processor Time	00:00:00.02
	Elapsed Time	00:00:00.01

Statistics

		Video 1 score	Video 2 score	Video 3 score	Video 4 score	Video 5 score	Video 6 score	Video 7 score	Video 8 score	Video 9 score
N	Valid	150	150	150	150	150	150	150	150	150
	Missing	0	0	0	0	0	0	0	0	0

Frequency Table

Video 1 score

	N	%
The patient passed the water swallow challenge.	150	100.0%

Video 2 score

	N	%
The patient passed the water swallow challenge.	1	0.7%
The patient failed the water swallow challenge.	147	98.0%
I'm not sure.	2	1.3%

Video 3 score

	N	%
The patient passed the water swallow challenge.	148	98.7%
The patient failed the water swallow challenge.	2	1.3%

Video 4 score

	N	%

The patient passed the water swallow challenge.	63	42.0%
The patient failed the water swallow challenge.	79	52.7%
I'm not sure.	8	5.3%

Video 5 score

	N	%
The patient passed the water swallow challenge.	17	11.3%
The patient failed the water swallow challenge.	127	84.7%
I'm not sure.	6	4.0%

Video 6 score

	N	%
The patient failed the water swallow challenge.	150	100.0%

Video 7 score

	N	%
The patient passed the water swallow challenge.	2	1.3%
The patient failed the water swallow challenge.	147	98.0%
I'm not sure.	1	0.7%

Video 8 score

	N	%
The patient passed the water swallow challenge.	11	7.3%
The patient failed the water swallow challenge.	135	90.0%
I'm not sure.	4	2.7%

Video 9 score

	N	%
The patient failed the water swallow challenge.	149	99.3%
I'm not sure.	1	0.7%

```
OUTPUT MODIFY
  /SELECT TABLES
```

```
  /IF COMMANDS=["Frequencies(LAST)"] SUBTYPES="Frequencies"
  /TABLECELLS SELECT=[VALIDPERCENT CUMULATIVEPERCENT]
APPLYTO=COLUMN HIDE=YES
  /TABLECELLS SELECT=[TOTAL] SELECTCONDITION=PARENT(VALID
MISSING) APPLYTO=ROW HIDE=YES
  /TABLECELLS SELECT=[VALID] APPLYTO=ROWHEADER UNGROUP=YES
  /TABLECELLS SELECT=[PERCENT] SELECTDIMENSION=COLUMNS
FORMAT="PCT" APPLYTO=COLUMN
  /TABLECELLS SELECT=[COUNT] APPLYTO=COLUMNHEADER
REPLACE="N"
  /TABLECELLS SELECT=[PERCENT] APPLYTO=COLUMNHEADER
REPLACE="%".
```

Interrater Reliability

Overall Agreement[a] for 9 effective subjects and 150 raters

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.697	.003	235.650	.000	.692	.703

Demographics

Coefficients[a]

		Collinearity Statistics	
Model		Tolerance	VIF
1	Dysphagia experience (years)	.615	1.627
	Dysphagia hours weekly	.517	1.933
	Dedicated swallowing course	.616	1.622
	Daily WST use	.542	1.844
	Weekly WST use	.641	1.560
	Monthly WST use	.814	1.229
	Annual WST use	.836	1.196
	Residential practice setting	.842	1.188
	Nonresidential practice setting	.745	1.342
	Community practice setting	.844	1.185
	School setting	.618	1.617
	Telehealth Setting	.822	1.217

	Terminal degree in progress	.945	1.058
	Terminal degree completed	.874	1.144

a. Dependent Variable: Rating task performance

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	6.554	15	.969
	Block	6.554	15	.969
	Model	6.554	15	.969

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.111	.055	1.836	.066
		Percent accuracy (original scoring) Dependent	.294	.140	1.836	.066
		Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP) Dependent	.069	.037	1.836	.066

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

	Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance

Ordinal by Ordinal	Somers' d	Symmetric	.025	.077	.330	.741
		Percent accuracy (original scoring) Dependent	.029	.088	.330	.741
		Modified Barium Swallow Impairment Profile (MBSImP) Dependent	.023	.068	.330	.741

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	-.059	.028	-1.659	.097
		Percent accuracy (original scoring) Dependent	-.297	.127	-1.659	.097
		Board Certified Specialist in Swallowing and Swallowing Disorders (BCS-S) Dependent	-.033	.020	-1.659	.097

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.069	.069	.993	.321
		Percent accuracy (original scoring) Dependent	.102	.102	.993	.321
		McNeill Dysphagia Therapy Program (MDTP) Dependent	.052	.052	.993	.321

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

T

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.013	.083	.159	.873
		Percent accuracy (original scoring) Dependent	.018	.112	.159	.873
		VitalStim Dependent	.011	.066	.159	.873

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.084	.071	1.144	.252
		Percent accuracy (original scoring) Dependent	.177	.150	1.144	.252
		Ampcare ESP Dependent	.055	.048	1.144	.252

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.031	.081	.383	.702
		Percent accuracy (original scoring) Dependent	.040	.104	.383	.702
		Lee Silverman Voice Therapy (LSVT) Dependent	.025	.066	.383	.702

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.015	.080	.194	.846
		Percent accuracy (original scoring) Dependent	.029	.150	.194	.846
		SPEAK OUT! Dependent	.011	.054	.194	.846

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Somers' *d* was run to determine the association between percent accuracy and DPNS

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.081	.065	1.188	.235
		Percent accuracy (original scoring) Dependent	.251	.197	1.188	.235
		Deep pharyngeal neuromuscular stimulation (DPNS) Dependent	.049	.041	1.188	.235

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	-.078	.077	-1.016	.310
		Percent accuracy (original scoring) Dependent	-.086	.085	-1.016	.310
		Flexible Endoscopic Evaluation of Swallowing (FEES) Dependent	-.071	.070	-1.016	.310

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.059	.078	.756	.450
		Percent accuracy (original scoring) Dependent	.065	.086	.756	.450
		Modified Barium Swallow Study (MBSS) Dependent	.054	.072	.756	.450

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	-.026	.011	-1.700	.089
		Percent accuracy (original scoring) Dependent	-.158	.054	-1.700	.089
		High resolution Pharyngeal Manometry (HRPM) Dependent	-.014	.008	-1.700	.089

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

```
GET
  FILE='\\files\users\amorrissey\Desktop\7.7 dataset.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
LOGISTIC REGRESSION VARIABLES Percent_category
  /METHOD=ENTER Yrs_Exp_Demographic Dys_week_Demographic
CCC_Demographic SwallCourse_Demographic
    WSTFreq_Demographic WSTFr_daily WSTFr_weekly WSTFr_monthly
WSTFr_yearly PracticeSetting_Hospital
    PracticeSetting_Residential PracticeSetting_Nonresidential
PracticeSetting_Community
    PracticeSetting_School PracticeSetting_Telehealth Edu_Masters_Degree
Edu_Term_in_process
    Edu_Term_completed Yrs_Exp_Demographic*ln_years
Dys_week_Demographic*ln_hours
  /CONTRAST (CCC_Demographic)=Indicator(1)
  /CONTRAST (SwallCourse_Demographic)=Indicator(1)
  /CONTRAST (WSTFreq_Demographic)=Indicator(1)
  /CONTRAST (WSTFr_daily)=Indicator(1)
  /CONTRAST (WSTFr_weekly)=Indicator(1)
  /CONTRAST (WSTFr_monthly)=Indicator(1)
  /CONTRAST (WSTFr_yearly)=Indicator(1)
  /CONTRAST (PracticeSetting_Hospital)=Indicator(1)
```

```
/CONTRAST (PracticeSetting_Residential)=Indicator(1)
/CONTRAST (PracticeSetting_Nonresidential)=Indicator(1)
/CONTRAST (PracticeSetting_Community)=Indicator(1)
/CONTRAST (PracticeSetting_School)=Indicator(1)
/CONTRAST (PracticeSetting_Telehealth)=Indicator(1)
/CONTRAST (Edu_Masters_Degree)=Indicator(1)
/CONTRAST (Edu_Term_in_process)=Indicator(1)
/CONTRAST (Edu_Term_completed)=Indicator(1)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Logistic Regression

Notes

Output Created		31-OCT-2022 15:31:30
Comments		
Input	Data	\\files\users\amorrissey\Desktop\7.7 dataset.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	150
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing

Syntax		LOGISTIC REGRESSION VARIABLES Percent_category /METHOD=ENTER Yrs_Exp_Demographic Dys_week_Demographic CCC_Demographic SwallCourse_Demographic WSTFreq_Demographic WSTFr_daily WSTFr_weekly WSTFr_monthly WSTFr_yearly PracticeSetting_Hospital PracticeSetting_Residential PracticeSetting_Nonresidential PracticeSetting_Community PracticeSetting_School PracticeSetting_Telehealth Edu_Masters_Degree Edu_Term_in_process Edu_Term_completed Yrs_Exp_Demographic*ln_years Dys_week_Demographic*ln_hours /CONTRAST (CCC_Demographic)=Indicator(1) /CONTRAST (SwallCourse_Demographic)=Indicator(1) /CONTRAST (WSTFreq_Demographic)=Indicator(1) /CONTRAST (WSTFr_daily)=Indicator(1) /CONTRAST (WSTFr_weekly)=Indicator(1) /CONTRAST (WSTFr_monthly)=Indicator(1) /CONTRAST (WSTFr_yearly)=Indicator(1) /CONTRAST (PracticeSetting_Hospital)=Indicator(1) /CONTRAST (PracticeSetting_Residential)=Indicator(1) /CONTRAST (PracticeSetting_Nonresidential)=Indicator(1) /CONTRAST (PracticeSetting_Community)=Indicator(1) /CONTRAST (PracticeSetting_School)=Indicator(1) /CONTRAST (PracticeSetting_Telehealth)=Indicator(1) /CONTRAST (Edu_Masters_Degree)=Indicator(1) /CONTRAST (Edu_Term_in_process)=Indicator(1) /CONTRAST (Edu_Term_completed)=Indicator(1) /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
Resources	Processor Time	00:00:00.02
	Elapsed Time	00:00:00.03

[DataSet1] \\files\users\amorrissey\Desktop\7.7 dataset.sav

Warnings

Due to redundancies, degrees of freedom have been reduced for one or more variables.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	150	100.0
	Missing Cases	0	.0
	Total	150	100.0
Unselected Cases		0	.0
Total		150	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
Below expert mean	0
Above expert mean	1

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Water swallow screening frequency	I don't use this.	53	.000	.000	.000	.000
	I use this daily.	38	1.000	.000	.000	.000
	I use this more than weekly but not daily.	32	.000	1.000	.000	.000
	I use this monthly but not weekly.	17	.000	.000	1.000	.000
	I use this a few times per year.	10	.000	.000	.000	1.000
Terminal degree completed	None	134	.000			
	Terminal degree completed.	16	1.000			
Dedicated swallowing course	No.	36	.000			
	Yes.	114	1.000			

I use this daily	None	112	.000
	I use this daily	38	1.000
More than weekly but not daily	None	118	.000
	I use this weekly but not daily	32	1.000
Monthly but not weekly.	None	133	.000
	Monthly but not weekly	17	1.000
Few times per year.	None	140	.000
	A few times a year	10	1.000
Hospital setting	None	70	.000
	Hospital setting	80	1.000
Residential Healthcare Facility	None	134	.000
	Residential healthcare facility	16	1.000
Terminal degree in progress	None	148	.000
	Terminal degree in progress	2	1.000
Master's degree completed.	None	18	.000
	Master's degree completed	132	1.000
Tele-health setting	None	148	.000
	Tele-health setting	2	1.000
School setting	None	131	.000
	School setting	19	1.000
Non-Residential Healthcare Setting	None	126	.000
	Non-residential healthcare setting	24	1.000
Community dwelling	None	141	.000
	Community dwelling	9	1.000
Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)	None	11	.000
	Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)	139	1.000

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Percent accuracy compared to experts		
	Observed		Below expert mean	Above expert mean	Percentage Correct
Step 0	Percent accuracy compared to experts	Below expert mean	0	21	.0
		Above expert mean	0	129	100.0
	Overall Percentage				86.0

a. Constant is included in the model.
b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.815	.235	59.513	1	<.001	6.143

Variables not in the Equation[a]

			Score	df	Sig.
Step 0	Variables	Dysphagia experience (years)	.068	1	.794
		Dysphagia hours weekly	.987	1	.321
		Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)(1)	1.737	1	.188
		Dedicated swallowing course(1)	1.166	1	.280
		Water swallow screening frequency	1.201	4	.878
		Water swallow screening frequency(1)	.510	1	.475
		Water swallow screening frequency(2)	.076	1	.783

	Water swallow screening frequency(3)	.080	1	.778
	Water swallow screening frequency(4)	.320	1	.571
	I use this daily(1)	.510	1	.475
	More than weekly but not daily(1)	.076	1	.783
	Monthly but not weekly.(1)	.080	1	.778
	Few times per year.(1)	.320	1	.571
	Hospital setting(1)	.142	1	.706
	Residential Healthcare Facility(1)	.893	1	.345
	Non-Residential Healthcare Setting(1)	.053	1	.817
	Community dwelling(1)	.066	1	.797
	School setting(1)	.058	1	.810
	Tele-health setting(1)	2.182	1	.140
	Master's degree completed.(1)	.121	1	.728
	Terminal degree in progress(1)	.330	1	.566
	Terminal degree completed(1)	.336	1	.562
	Dysphagia experience (years) by Natural Log Transformation of "Years"	.016	1	.900
	Dysphagia hours weekly by Natural Log Transformation of "Hours"	1.110	1	.292

a. Residual Chi-Squares are not computed because of redundancies.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.149	17	.935
	Block	9.149	17	.935

	Model	9.149	17	.935

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	112.340[a]	.059	.107

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

			Predicted		
			Percent accuracy compared to experts		
	Observed		Below expert mean	Above expert mean	Percentage Correct
Step 1	Percent accuracy compared to experts	Below expert mean	1	20	4.8
		Above expert mean	0	129	100.0
	Overall Percentage				86.7

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Dysphagia experience (years)	.327	.257	1.620	1	.203	1.387
	Dysphagia hours weekly	-.323	.242	1.786	1	.181	.724
	Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)(1)	.306	.912	.113	1	.737	1.358
	Dedicated swallowing course(1)	.711	.704	1.019	1	.313	2.035

Water swallow screening frequency			1.164	4	.884	
Water swallow screening frequency(1)	.739	.817	.817	1	.366	2.093
Water swallow screening frequency(2)	.361	.790	.208	1	.648	1.434
Water swallow screening frequency(3)	.564	.942	.359	1	.549	1.758
Water swallow screening frequency(4)	-.304	.974	.097	1	.755	.738
Hospital setting(1)	1.028	1.806	.324	1	.569	2.797
Residential Healthcare Facility(1)	1.894	2.031	.870	1	.351	6.648
Non-Residential Healthcare Setting(1)	1.297	1.816	.510	1	.475	3.658
Community dwelling(1)	2.052	1.944	1.114	1	.291	7.783
School setting(1)	1.353	1.658	.666	1	.414	3.870
Master's degree completed.(1)	.247	.758	.106	1	.745	1.280
Terminal degree in progress(1)	19.390	27981.910	.000	1	.999	263643418.229
Dysphagia experience (years) by Natural Log Transformation of "Years"	-.082	.068	1.444	1	.230	.921
Dysphagia hours weekly by Natural Log Transformation of "Hours"	.199	.143	1.928	1	.165	1.220
Constant	-.908	1.784	.259	1	.611	.403

a. Variable(s) entered on step 1: Dysphagia experience (years), Dysphagia hours weekly, Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP), Dedicated swallowing course, Water swallow screening frequency, Hospital setting, Residential Healthcare Facility, Non-Residential Healthcare Setting, Community dwelling, School setting, Master's degree completed., Terminal degree in progress, Dysphagia experience (years) * Natural Log Transformation of "Years" , Dysphagia hours weekly * Natural Log Transformation of "Hours" .

```
LOGISTIC REGRESSION VARIABLES Percent_category
  /METHOD=ENTER Yrs_Exp_Demographic Dys_week_Demographic
CCC_Demographic SwallCourse_Demographic
    WSTFreq_Demographic WSTFr_daily WSTFr_weekly WSTFr_monthly
WSTFr_yearly PracticeSetting_Hospital
    PracticeSetting_Residential PracticeSetting_Nonresidential
PracticeSetting_Community
    PracticeSetting_School PracticeSetting_Telehealth Edu_Masters_Degree
Edu_Term_in_process
    Edu_Term_completed
  /CONTRAST (CCC_Demographic)=Indicator(1)
  /CONTRAST (SwallCourse_Demographic)=Indicator(1)
  /CONTRAST (WSTFreq_Demographic)=Indicator(1)
  /CONTRAST (WSTFr_daily)=Indicator(1)
  /CONTRAST (WSTFr_weekly)=Indicator(1)
  /CONTRAST (WSTFr_monthly)=Indicator(1)
  /CONTRAST (WSTFr_yearly)=Indicator(1)
  /CONTRAST (PracticeSetting_Hospital)=Indicator(1)
  /CONTRAST (PracticeSetting_Residential)=Indicator(1)
  /CONTRAST (PracticeSetting_Nonresidential)=Indicator(1)
  /CONTRAST (PracticeSetting_Community)=Indicator(1)
  /CONTRAST (PracticeSetting_School)=Indicator(1)
  /CONTRAST (PracticeSetting_Telehealth)=Indicator(1)
  /CONTRAST (Edu_Masters_Degree)=Indicator(1)
  /CONTRAST (Edu_Term_in_process)=Indicator(1)
  /CONTRAST (Edu_Term_completed)=Indicator(1)
  /SAVE=PRED
  /CLASSPLOT
  /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT SUMMARY CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Logistic Regression

Notes

Output Created		31-OCT-2022 16:13:32
Comments		
Input	Data	\\files\users\amorrissey\Desktop\7.7 dataset.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	150
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing

Syntax

```
LOGISTIC REGRESSION VARIABLES
Percent_category
/METHOD=ENTER
Yrs_Exp_Demographic
Dys_week_Demographic CCC_Demographic
SwallCourse_Demographic
WSTFreq_Demographic WSTFr_daily
WSTFr_weekly WSTFr_monthly
WSTFr_yearly PracticeSetting_Hospital
PracticeSetting_Residential
PracticeSetting_Nonresidential
PracticeSetting_Community
PracticeSetting_School
PracticeSetting_Telehealth
Edu_Masters_Degree Edu_Term_in_process
Edu_Term_completed
/CONTRAST
(CCC_Demographic)=Indicator(1)
/CONTRAST
(SwallCourse_Demographic)=Indicator(1)
/CONTRAST
(WSTFreq_Demographic)=Indicator(1)
/CONTRAST (WSTFr_daily)=Indicator(1)
/CONTRAST
(WSTFr_weekly)=Indicator(1)
/CONTRAST
(WSTFr_monthly)=Indicator(1)
/CONTRAST (WSTFr_yearly)=Indicator(1)
/CONTRAST
(PracticeSetting_Hospital)=Indicator(1)
/CONTRAST
(PracticeSetting_Residential)=Indicator(1)
/CONTRAST
(PracticeSetting_Nonresidential)=Indicator(1)
/CONTRAST
(PracticeSetting_Community)=Indicator(1)
/CONTRAST
(PracticeSetting_School)=Indicator(1)
```

		/CONTRAST (PracticeSetting_Telehealth)=Indicator(1) /CONTRAST (Edu_Masters_Degree)=Indicator(1) /CONTRAST (Edu_Term_in_process)=Indicator(1) /CONTRAST (Edu_Term_completed)=Indicator(1) /SAVE=PRED /CLASSPLOT /CASEWISE OUTLIER(2) /PRINT=GOODFIT SUMMARY CI(95) /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
Resources	Processor Time	00:00:00.02
	Elapsed Time	00:00:00.03
Variables Created or Modified	PRE_6	Predicted probability

Warnings

Due to redundancies, degrees of freedom have been reduced for one or more variables.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	150	100.0
	Missing Cases	0	.0
	Total	150	100.0
Unselected Cases		0	.0
Total		150	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
Below expert mean	0
Above expert mean	1

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Water swallow screening frequency	I don't use this.	53	.000	.000	.000	.000
	I use this daily.	38	1.000	.000	.000	.000
	I use this more than weekly but not daily.	32	.000	1.000	.000	.000
	I use this monthly but not weekly.	17	.000	.000	1.000	.000
	I use this a few times per year.	10	.000	.000	.000	1.000
Terminal degree completed	None	134	.000			
	Terminal degree completed.	16	1.000			
Dedicated swallowing course	No.	36	.000			
	Yes.	114	1.000			
I use this daily	None	112	.000			
	I use this daily	38	1.000			
More than weekly but not daily	None	118	.000			
	I use this weekly but not daily	32	1.000			
Monthly but not weekly.	None	133	.000			
	Monthly but not weekly	17	1.000			
Few times per year.	None	140	.000			
	A few times a year	10	1.000			
Hospital setting	None	70	.000			
	Hospital setting	80	1.000			
Residential Healthcare Facility	None	134	.000			
	Residential healthcare facility	16	1.000			
Terminal degree in progress	None	148	.000			
	Terminal degree in progress	2	1.000			
Master's degree completed.	None	18	.000			
	Master's degree completed	132	1.000			

Tele-health setting	None	148	.000
	Tele-health setting	2	1.000
School setting	None	131	.000
	School setting	19	1.000
Non-Residential Healthcare Setting	None	126	.000
	Non-residential healthcare setting	24	1.000
Community dwelling	None	141	.000
	Community dwelling	9	1.000
Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)	None	11	.000
	Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)	139	1.000

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Percent accuracy compared to experts		
	Observed		Below expert mean	Above expert mean	Percentage Correct
Step 0	Percent accuracy compared to experts	Below expert mean	0	21	.0
		Above expert mean	0	129	100.0
	Overall Percentage				86.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.815	.235	59.513	1	<.001	6.143

Variables not in the Equation[a]

			Score	df	Sig.
Step 0	Variables	Dysphagia experience (years)	.068	1	.794
		Dysphagia hours weekly	.987	1	.321
		Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)(1)	1.737	1	.188
		Dedicated swallowing course(1)	1.166	1	.280
		Water swallow screening frequency	1.201	4	.878
		Water swallow screening frequency(1)	.510	1	.475
		Water swallow screening frequency(2)	.076	1	.783
		Water swallow screening frequency(3)	.080	1	.778
		Water swallow screening frequency(4)	.320	1	.571
		I use this daily(1)	.510	1	.475
		More than weekly but not daily(1)	.076	1	.783
		Monthly but not weekly.(1)	.080	1	.778
		Few times per year.(1)	.320	1	.571
		Hospital setting(1)	.142	1	.706
		Residential Healthcare Facility(1)	.893	1	.345
		Non-Residential Healthcare Setting(1)	.053	1	.817
		Community dwelling(1)	.066	1	.797
		School setting(1)	.058	1	.810
		Tele-health setting(1)	2.182	1	.140
		Master's degree completed.(1)	.121	1	.728

	Terminal degree in progress(1)	.330	1	.566
	Terminal degree completed(1)	.336	1	.562

a. Residual Chi-Squares are not computed because of redundancies.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	6.554	15	.969
	Block	6.554	15	.969
	Model	6.554	15	.969

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	114.935[a]	.043	.077

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.394	8	.310

Contingency Table for Hosmer and Lemeshow Test

		Percent accuracy compared to experts = Below expert mean		Percent accuracy compared to experts = Above expert mean		
		Observed	Expected	Observed	Expected	Total
Step 1	1	5	4.638	10	10.362	15
	2	2	3.010	13	11.990	15
	3	4	2.572	11	12.428	15
	4	2	2.154	13	12.846	15
	5	1	1.966	14	13.034	15
	6	4	1.783	11	13.217	15

	7	0	1.597	15	13.403	15
	8	0	1.478	16	14.522	16
	9	2	1.162	13	13.838	15
	10	1	.640	13	13.360	14

Classification Table[a]

			Predicted		
			Percent accuracy compared to experts		
	Observed		Below expert mean	Above expert mean	Percentage Correct
Step 1	Percent accuracy compared to experts	Below expert mean	1	20	4.8
		Above expert mean	0	129	100.0
	Overall Percentage				86.7

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Dysphagia experience (years)	.013	.031	.177	1	.674	1.013	.954	1.076
	Dysphagia hours weekly	.011	.027	.183	1	.669	1.011	.960	1.066
	Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)(1)	.759	.845	.807	1	.369	2.135	.408	11.179

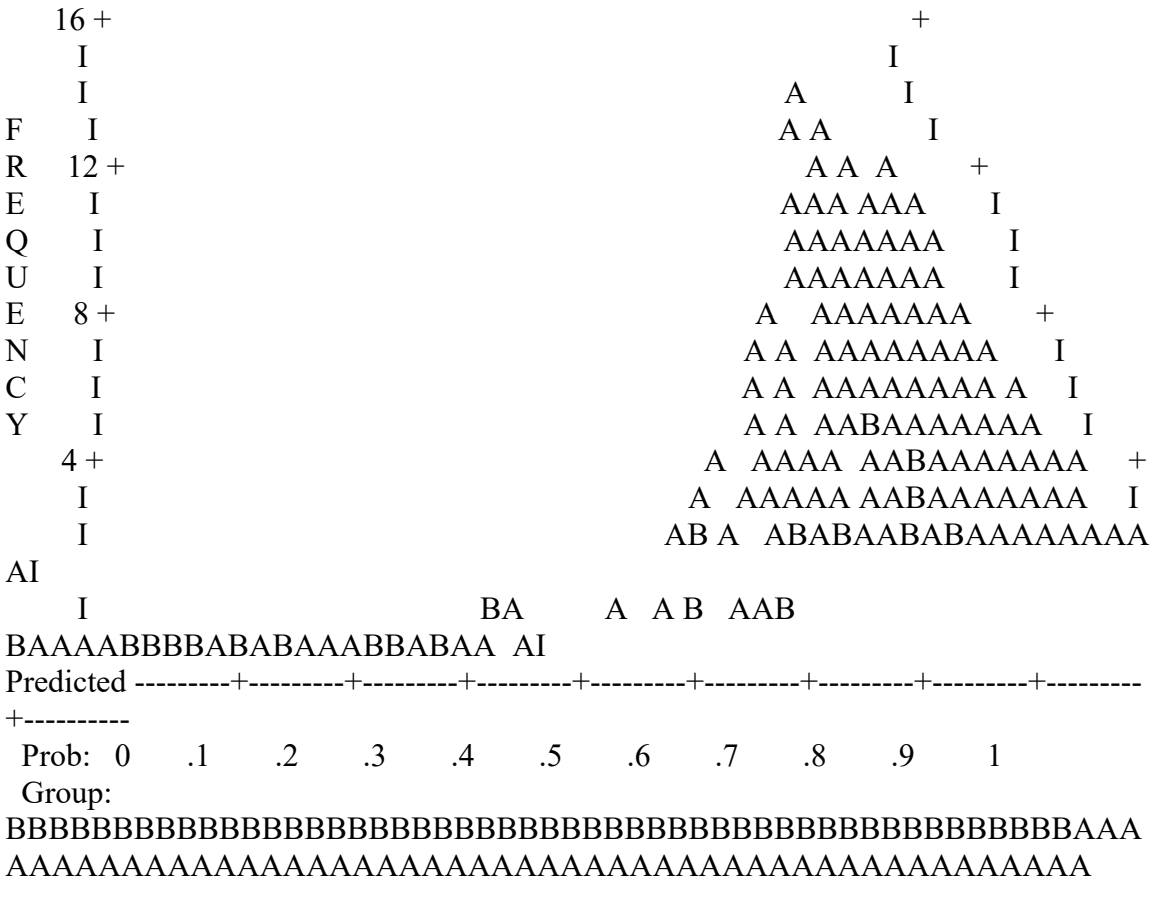
	Dedicated swallowing course(1)	.584	.686	.724	1	.395	1.793	.467	6.886
	Water swallow screening frequency			.482	4	.975			
	Water swallow screening frequency(1)	.430	.796	.292	1	.589	1.537	.323	7.310
	Water swallow screening frequency(2)	-.010	.732	.000	1	.989	.990	.236	4.162
	Water swallow screening frequency(3)	.336	.888	.143	1	.705	1.400	.246	7.973
	Water swallow screening frequency(4)	-.068	.950	.005	1	.943	.934	.145	6.012
	Hospital setting(1)	.841	1.721	.239	1	.625	2.318	.079	67.636
	Residential Healthcare Facility(1)	1.806	1.953	.856	1	.355	6.088	.132	279.774
	Non-Residential Healthcare Setting(1)	1.217	1.776	.469	1	.493	3.376	.104	109.733
	Community dwelling(1)	1.506	1.880	.642	1	.423	4.509	.113	179.456
	School setting(1)	1.369	1.625	.710	1	.399	3.933	.163	95.080

	Master's degree completed.(1)	.223	.757	.087	1	.769	1.249	.283	5.506
	Terminal degree in progress(1)	19.572	28409.738	.000	1	.999	316098515.482	.000	.
	Constant	-1.001	1.753	.326	1	.568	.368		

a. Variable(s) entered on step 1: Dysphagia experience (years), Dysphagia hours weekly, Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP), Dedicated swallowing course, Water swallow screening frequency, Hospital setting, Residential Healthcare Facility, Non-Residential Healthcare Setting, Community dwelling, School setting, Master's degree completed., Terminal degree in progress.

Step number: 1

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for Above expert mean
The Cut Value is .50
Symbols: B - Below expert mean
A - Above expert mean
Each Symbol Represents 1 Case.

Casewise List[b]

		Observed			Temporary Variable		
Case	Selected Status[a]	Percent accuracy compared to experts	Predicted	Predicted Group	Resid	ZResid	SResid
2	S	B**	.874	A	-.874	-2.630	-2.097
3	S	B**	.857	A	-.857	-2.448	-2.020
4	S	B**	.812	A	-.812	-2.077	-2.006
7	S	B**	.877	A	-.877	-2.669	-2.305
8	S	B**	.878	A	-.878	-2.680	-2.118
11	S	B**	.860	A	-.860	-2.478	-2.018
13	S	B**	.926	A	-.926	-3.535	-2.399
14	S	B**	.878	A	-.878	-2.685	-2.131
15	S	B**	.919	A	-.919	-3.358	-2.310
16	S	B**	.877	A	-.877	-2.675	-2.115
19	S	B**	.949	A	-.949	-4.336	-2.536

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

```
SORT CASES BY RaterID (A).
ROC PRE_1 BY Percent_category (1)
  /PLOT=CURVE(REFERENCE)
  /PRINT=SE
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE)
CI(95)
  /MISSING=EXCLUDE.
```

ROC Curve

Notes

Output Created	31-OCT-2022 18:10:14
Comments	

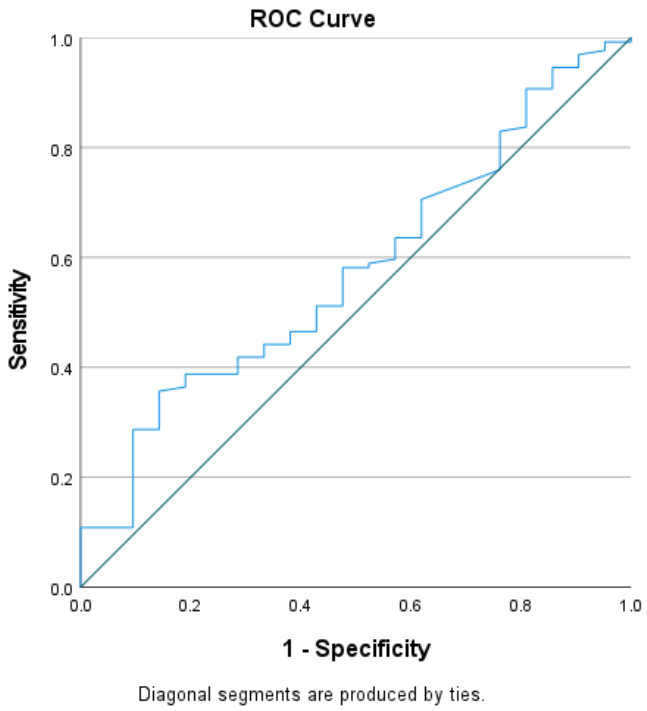
Input	Data	\\files\users\amorrissey\Desktop\7.7 dataset.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	150
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data for all variables in the analysis.
Syntax		ROC PRE_1 BY Percent_category (1) /PLOT=CURVE(REFERENCE) /PRINT=SE /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95) /MISSING=EXCLUDE.
Resources	Processor Time	00:00:03.33
	Elapsed Time	00:00:16.45

Case Processing Summary

Percent accuracy compared to experts	Valid N (listwise)
Positive[a]	129
Negative	21

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is Above expert mean.



Area Under the Curve

Test Result Variable(s): Predicted probability

			Asymptotic 95% Confidence Interval	
Area	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
.578	.063	.253	.454	.702

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

```
GET
  FILE='\\files\users\amorrissey\Desktop\7.7 dataset.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	114.935[a]	.043	.077

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

			Predicted		
			Percent accuracy compared to experts		
	Observed		Below expert mean	Above expert mean	Percentage Correct
Step 1	Percent accuracy compared to experts	Below expert mean	1	20	4.8
		Above expert mean	0	129	100.0
	Overall Percentage				86.7

a. The cut value is .500

_Area Under the Curve

Test Result Variable(s): Predicted probability

			Asymptotic 95% Confidence Interval	
Area	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
.587	.063	.253	.454	.702

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	150	100.0
	Missing Cases	0	.0
	Total	150	100.0
Unselected Cases		0	.0
Total		150	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
Below expert mean	0
Above expert mean	1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Dysphagia experience (years)	.013	.031	.177	1	.674	1.013	.954	1.076
	Dysphagia hours weekly	.011	.027	.183	1	.669	1.011	.960	1.066
	Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP)(1)	.759	.845	.807	1	.369	2.135	.408	11.179
	Dedicated swallowing course(1)	.584	.686	.724	1	.395	1.793	.467	6.886
	Water swallow screening use			.482	4	.975			
	Water swallow screening frequency (daily)	.430	.796	.292	1	.589	1.537	.323	7.310
	Water swallow screening frequency (weekly)	-.010	.732	.000	1	.989	.990	.236	4.162
	Water swallow screening frequency (monthly)	.336	.888	.143	1	.705	1.400	.246	7.973
	Water swallow screening frequency (annually)	-.068	.950	.005	1	.943	.934	.145	6.012
	Hospital setting(1)	.841	1.721	.239	1	.625	2.318	.079	67.636

Residential Healthcare Facility(1)	1.806	1.953	.856	1	.355	6.088	.132	279.774
Non-Residential Healthcare Setting(1)	1.217	1.776	.469	1	.493	3.376	.104	109.733
Community dwelling(1)	1.506	1.880	.642	1	.423	4.509	.113	179.456
School setting(1)	1.369	1.625	.710	1	.399	3.933	.163	95.080
Master's degree completed.(1)	.223	.757	.087	1	.769	1.249	.283	5.506
Terminal degree in progress(1)	19.572	28409.738	.000	1	.999	3160985 15.482	.000	.
Constant	-1.001	1.753	.326	1	.568	.368		

a. Variable(s) entered on step 1: Dysphagia experience (years), Dysphagia hours weekly, Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP), Dedicated swallowing course, Water swallow screening frequency, Hospital setting, Residential Healthcare Facility, Non-Residential Healthcare Setting, Community dwelling, School setting, Master's degree completed., Terminal degree in progress.

Final Study: Percent correct and proportion of agreement for each rater

Rater	% Correct	Difference of CSV ratings (initial, repeated)	Rater Proportion of Agreement
1	77.77	0	100
2	88.88	0	100
3	88.88	0	100
4	88.88	0	100
5	88.88	0	100
6	88.88	0	100
7	88.88	1/2	50
8	88.88	0	100
9	88.88	0	100
10	88.88	0	100
11	88.88	0	100
12	88.88	0	100
13	88.88	0	100
14	88.88	0	100
15	88.88	0	100
16	88.88	0	100
17	77.77	0	100
18	77.77	0	100
19	88.88	0	100
20	88.88	0	100
21	88.88	0	100
22	88.88	0	100
23	77.77	0	100
24	88.88	0	100
25	88.88	0	100
26	88.88	0	100
27	88.88	0	100
28	99.99	0	100
29	77.77	1/2	50
30	77.77	1/2	50
31	88.88	0	100
32	88.88	0	100
33	88.88	0	100
34	88.88	0	100
35	88.88	0	100
36	88.88	0	100
37	77.77	0	100
38	88.88	0	100
39	88.88	0	100
40	77.77	0	100

41	99.99	0	100
42	99.99	0	100
43	88.88	0	100
44	88.88	0	100
45	99.99	0	100
46	99.99	0	100
47	88.88	0	100
48	88.88	0	100
49	99.99	0	100
50	88.88	0	100
51	88.88	0	100
52	88.88	0	100
53	99.99	0	100
54	88.88	0	100
55	88.88	0	100
56	99.99	0	100
57	88.88	0	100
58	88.88	0	100
59	88.88	0	100
60	88.88	0	100
61	99.99	0	100
62	99.99	0	100
63	99.99	0	100
64	99.99	0	100
65	88.88	0	100
66	99.99	0	100
67	88.88	0	100
68	66.66	0	100
69	77.77	0	100
70	88.88	0	100
71	99.99	0	100
72	88.88	0	100
73	88.88	1/2	50
74	99.99	0	100
75	66.66	0	100
76	88.88	0	100
77	99.99	0	100
78	88.88	0	100
79	88.88	0	100
80	66.66	0	100
81	99.99	0	100
82	99.99	0	100
83	88.88	0	100
84	99.99	0	100
85	99.99	0	100

86	88.88	0	100
87	99.99	0	100
88	99.99	0	100
89	88.88	0	100
90	88.88	0	100
91	99.99	0	100
92	88.88	0	100
93	88.88	0	100
94	99.99	0	100
95	88.88	0	100
96	77.77	0	100
97	88.88	0	100
98	88.88	1/2	50
99	99.99	0	100
100	55.55	0	100
101	66.66	1/2	50
102	99.99	0	100
103	77.77	0	100
104	88.88	1/2	50
105	99.99	0	100
106	88.88	1/2	50
107	99.99	0	100
108	77.77	1/2	50
109	88.88	0	100
110	88.88	0	100
111	88.88	0	100
112	88.88	0	100
113	88.88	0	100
114	77.77	0	100
115	99.99	0	100
116	88.88	1/2	50
117	99.99	0	100
118	88.88	0	100
119	88.88	0	100
120	99.99	0	100
121	88.88	0	100
122	88.88	0	100
123	99.99	0	100
124	99.99	0	100
125	88.88	0	100
126	88.88	0	100
127	88.88	0	100
128	77.77	0	100
129	99.99	0	100
130	99.99	0	100

131	99.99	0	100
132	99.99	0	100
133	88.88	0	100
134	66.66	0	100
135	88.88	0	100
136	88.88	1/2	50
137	77.77	1/2	50
138	88.88	1/2	50
139	99.99	0	100
140	88.88	0	100
141	99.99	0	100
142	99.99	0	100
143	88.88	0	100
144	88.88	0	100
145	88.88	0	100
146	99.99	0	100
147	99.99	0	100
148	88.88	0	100
149	99.99	0	100
150	99.99	0	100

Note. Difference score of 0 indicates consistent judgments between initial and repeated CSVs ratings. Difference score of ½ indicates consistent judgments of one of two CSVs.

Individual Response Category Output: Pilot & Final Study

Pilot Study

Individual kappa for the "*The patient passed the 3-ounce water swallow challenge"* (0) and *"The patient failed the 3-ounce water swallow challenge"* (1) were obtained. The *I'm not sure"* judgement did not occur during cup drinking.

Although Fleiss' Kappa was used to determine level of agreement as mentioned above, Cohen's Kappa scale was used to interpret strength. These results indicate that raters demonstrated 'good' agreement for 3-ounce water swallow challenge videos presented via cup delivery method when ratings of "*The patient passed the 3-ounce water swallow challenge*" and "*The patient failed the 3-ounce water swallow challenge*" were assigned (Table _).

Table _
Cup Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	.803	**.721**	.063	11.453	.000	**.598**	**.845**
1	.919	**.721**	.063	11.453	.000	**.598**	**.845**

a. Sample data contains 9 effective subjects and 8 raters.

Using Cohen's Kappa scale for the straw delivery method, there was moderate agreement between raters and a statistically significant result, kappa = .514, 95% CI [.413, .616], p =.000 (see Table _). Fleiss' Kappa was statistically significantly different to zero (p=.000), rejecting the null hypothesis and failing to reject the alternative hypothesis.

Individual kappa for "*The patient passed the 3-ounce water swallow challenge"* (0), *"The patient failed the 3-ounce water swallow challenge"* (1), and "*I'm not sure"* (2) were assigned. Although Fleiss' Kappa was used to determine level of agreement, Cohen's Kappa scale was used to interpret strength. These results indicate that raters demonstrated 'fair' agreement for 3-ounce water swallow challenge videos presented via straw delivery method when ratings of "*The patient passed the 3-ounce water swallow challenge*" (0) and *I'm not sure"* (2) were assigned. There was 'good' agreement for 3-ounce water swallow challenge videos receiving a "*The patient failed the 3-ounce water swallow challenge*" designation.

Table _

Straw Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	.367	**.290**	.067	4.335	<.001	.159	.421
1	.956	**.766**	.067	11.458	.000	.635	.897
2	.343	**.287**	.067	4.298	<.001	.156	.418

b. Sample data contains 8 effective subjects and 8 raters.

Final Study

An individual Kappa for each response category was completed. "*The patient passed the 3-ounce water swallow challenge"* (0) response category revealed "good" strength of agreement (κ = .697, 95% CI [0.69, 0.70], $p < .00$). *"The patient failed the 3-ounce water swallow challenge"* (1) response category revealed "good" strength of agreement, (κ = .70, 95% CI [0.70, 0.71], $p < .00$). The *"I'm not sure"* (2) response category revealed 'poor' strength of agreement (κ = .013, 95% CI [0.01, 0.02], $p < .001$).

These results indicated that participants demonstrated 'good' agreement for CSVs when ratings of "*The patient passed the 3-ounce water swallow challenge*" and "*The patient failed the 3-ounce water swallow challenge*" were assigned. There was 'poor' agreement for CSVs which received the *"I'm not sure"* designation. This confirms a high interrater agreement overall for CSVs across videos. Participants also demonstrated a high level of agreement when assigning "*The patient passed the 3-ounce water swallow challenge" and* "*The patient failed the 3-ounce water swallow challenge."* Participants revealed markedly low agreement for response category 2, *"I'm not sure."* This was attributed to the very low frequency with which this response occurred.

Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	.819	.744	.003	237.477	.000	.738	.750
1	.909	.706	.003	225.291	.000	.700	.712
2	.030	.013	.003	4.293	<.001	.007	.020

a. Sample data contains 9 effective subjects and 151 raters.

Individual kappa for the "*The patient passed the 3-ounce water swallow challenge"* (0), *"The patient failed the 3-ounce water swallow challenge"* (1), and *"I'm not sure"* (2) categories were .615, .743, and .318, respectively (Table 12). Although Fleiss' Kappa was used to determine level of agreement as mentioned above, Cohen's Kappa scale was used to interpret strength. These results indicate that raters demonstrated 'good 'agreement for 3-ounce water swallow challenge videos when ratings of "*The patient passed the 3-ounce water swallow challenge*" and "*The patient failed the 3-ounce water swallow challenge*" were assigned. There was 'fair' agreement for 3-ounce water swallow challenge videos receiving the *"I'm not sure"* designation.