Dysphagia

Psychometric Properties of Visuoperceptual Measures of Videofluoroscopic and Fibre-Endoscopic Evaluations of Swallowing: A Systematic Review

--Manuscript Draft--

Manuscript Number:	
Full Title:	Psychometric Properties of Visuoperceptual Measures of Videofluoroscopic and Fibre-Endoscopic Evaluations of Swallowing: A Systematic Review
Article Type:	Invited Reviews and Submitted Reviews
Keywords:	Videofluoroscopy; Fibre-Endoscopic Evaluations of Swallowing; Dysphagia; Deglutition; Measure; Psychometrics.
Corresponding Author:	Katina Swan, BSpPath (hons) Curtin University Bentley, Western Australia AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Curtin University
Corresponding Author's Secondary Institution:	
First Author:	Katina Swan, BSpPath (hons)
First Author Secondary Information:	
Order of Authors:	Katina Swan, BSpPath (hons)
	Reinie Cordier, PhD
	Ted Brown, PhD
	Renee Speyer, PhD
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Introduction Fibreoptic Endoscopic Evaluation of Swallowing (FEES) and Videofluoroscopic Swallow Studies (VFSS) are instrumental assessments utilised in dysphagia which provide real-time videos of the internal structures of swallowing. They are commonly regarded as 'gold-standard' assessments; however, there is no consensus regarding a gold-standard measure to analyse the video recordings they produce. Measures require sound psychometric properties to be suitable for clinical or research purposes. To date, no review of psychometric properties of FEES and VFSS measures has been undertaken or formally reported. Objective This review assessed the quality of the psychometric properties of visuoperceptual measures of FEES and VFSS. Methods Electronic databases were searched for studies reporting on psychometric qualities of visuoperceptual measures which are used to analyse recordings from FEES and VFSS. All dates until February 2017 were included. The Consensus based Standards for the selection of health Measurement Instruments (COSMIN) checklist was used to evaluate methodical quality of studies. The measures' overall quality was then assessed by combining COSMIN ratings with quality criteria. Results Forty-five studies met inclusion criteria for this review, which reported details on 39 measures. Data about the measures' psychometric properties was very limited. Twenty-one measures had information available about reliability only, while 18 had information on two to four psychometric properties of the possible nine categorised within the COSMIN framework. The majority of the FEES and VFSS measures' psychometric properties were rated as 'indeterminate' overall, due to the small number

	of studies and issues with design, statistical analyses and reporting of extant studies. Conclusions There is insufficient evidence to recommend any individual measure included in this review as valid and reliable to interpret VFSS and FEES recordings. Further research is needed regarding psychometric properties of measures for FEES and VFSS, which utilises robust methodological design and reporting.
Suggested Reviewers:	Deborah Denman, BSpPath deborah.denman@postgrad.curtin.edu.au Ms Denman is a speech pathologist with experience using the COSMIN tool to analyse the psychometric prtoperties of studies. She is currently completing a higher degree by research.
	Hans Bogaardt, PhD hans.bogaardt@sydney.edu.au Dr Bogaardt is a Speech Pathologist and Clinical Epidemiologist, who is specialized in assessment and treatment of dysphagia. He has experience analysing the psychometric qualities of measures.
	Amy Hodges, BOccThpy amy.hodges@curtin.edu.au Ms Hodges has experience using the COSMIN checklist to analyse the psychometric qualities of measures and is currently completing a higher degree by research
	Jae Hyun-Kim, PhD Macquarie University jae-hyun.kim@mq.edu.au Dr Kim is a speech pathologist who has experience using the COSMIN checklist to analyse the psychometric qualities of measures.
	Daniele Farneti, MD, PhD dfarneti@auslrn.net; daniele.farneti@unibo.it Dr Farneti is an ENT with experience in dysphagia, FEES and measure development.
Opposed Reviewers:	

Running title: Systematic Review of Visuoperceptual Measures for Instrumental Assessments of Dysphagia

Psychometric Properties of Visuoperceptual Measures of Videofluoroscopic and Fibre-Endoscopic Evaluations of Swallowing: A Systematic Review*

Katina Swan, BSpPath(Hons)[1], Reinie Cordier, Ph.D[1], Ted Brown, Ph.D[2], Renée Speyer, Ph.D[21,3,4]

1. School of Occupational Therapy and Social Work, Curtin University, Perth, WA, Australia.
2. Department of Occupational Therapy, School of Primary and Allied Health Care, Faculty of Medicine, Nursing and Health Sciences, Monash University – Peninsula Campus, Frankston, VIC, Australia.
3. Department of Special Needs Education, University of Oslo, Oslo, Norway.
4. Department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Centre, Leiden, the Netherlands.

Corresponding Author:

Katina Swan

School of Occupational Therapy and Social Work, Curtin University, Perth, W.A., Australia

katina.swan@postgrad.curtin.edu.au

Reprint address:

A/ Prof Reinie Cordier

GPO Box U1987, Perth WA 6845

Tel: +61 8 9266 2583

Declaration of interest:

The authors have no competing interests to declare.

PROSPERO Registration No: CRD42017060032

*The first author completed this study as part of the requirements for the completion of a PhD under supervision of Reinie Cordier, Ted Brown and Renée Speyer. The authors wish to acknowledge Curtin University and the Australian Federal Government for the Curtin University Postgraduate Scholarship (CUPS) and the Australian Postgraduate Award (APA). The authors of the study would like to thank Ms Amy Hodges, who assisted with abstract screening and instrument ratings.

Running title: Systematic Review of Visuoperceptual Measures for Instrumental Assessments of Dysphagia

Psychometric Properties of Visuoperceptual Measures of Videofluoroscopic and Fibre-Endoscopic Evaluations of Swallowing: A Systematic Review

Abstract

Introduction

Fibreoptic Endoscopic Evaluation of Swallowing (FEES) and Videofluoroscopic Swallow Studies (VFSS) are instrumental assessments utilised in dysphagia which provide real-time videos of the internal structures of swallowing. They are commonly regarded as 'gold-standard' assessments; however, there is no consensus regarding a gold-standard measure to analyse the video recordings they produce. Measures require sound psychometric properties to be suitable for clinical or research purposes. To date, no review of psychometric properties of FEES and VFSS measures has been undertaken or formally reported.

Objective

This review assessed the quality of the psychometric properties of visuoperceptual measures of FEES and VFSS.

Methods

Electronic databases were searched for studies reporting on psychometric qualities of visuoperceptual measures which are used to analyse recordings from FEES and VFSS. All dates until February 2017 were included. The Consensus based Standards for the selection of health Measurement Instruments (COSMIN) checklist was used to evaluate methodical quality of studies. The measures' overall quality was then assessed by combining COSMIN ratings with quality criteria.

Results

Forty-five studies met inclusion criteria for this review, which reported details on 39 measures. Data about the measures' psychometric properties was very limited. Twenty-one measures had information available about reliability only, while 18 had information on two to four psychometric properties of the possible nine categorised within the COSMIN framework. The majority of the FEES and VFSS measures' psychometric properties were rated as 'indeterminate' overall, due to the small number of studies and issues with design, statistical analyses and reporting of extant studies.

Conclusions

There is insufficient evidence to recommend any individual measure included in this review as valid and reliable to interpret VFSS and FEES recordings. Further research is needed regarding

psychometric properties of measures for FEES and VFSS, which utilises robust methodological design and reporting.

Key Words:

Videofluoroscopy; Fibre-Endoscopic Evaluations of Swallowing; Dysphagia; Deglutition; Measure; Psychometrics.

Introduction

Dysphagia is associated with many common conditions, including premature birth, developmental disabilities, head and neck cancer, neurodegenerative diseases, acquired brain injury and stroke (2-5). It occurs across a range of settings and regions; in the Netherlands, prevalence in the general population has been reported to be as high as 12.1% (6). A British study reported up to 1 in 9 community-dwelling older adults are impacted by dysphagia (7), while South Korean research found an incidence of 52.7% among older adults in nursing homes (8). Up to 30% of acutely hospitalised patients may be affected by dysphagia (9) and nearly a quarter of infants who undergo open-heart surgery have dysphagia symptoms (10). In addition to malnutrition, dehydration and choking, dysphagia may also cause acute lung infection, known as aspiration pneumonia. Aspiration pneumonia is the result of material from the oral, pharyngeal or gastric regions entering the lungs (11) and is a strong independent predictor of mortality at 30 days post admission compared to community and hospital-acquired pneumonias. Among patients with aspiration pneumonia, median length of stay in hospital is increased by 8.5 days (12). Dysphagia has also been found to profoundly affect quality of life (13, 14). For example, difficulty swallowing can cause frustration, anxiety and embarrassment during mealtimes and special social events which should be pleasurable (15).

These issues underscore the need for high-quality assessment practices where dysphagia is concerned. Dysphagia assessment typically first takes place at the home, clinic or the bedside where clinicians gather patient history and concerns and use non-invasive testing to assess nervous and muscle function and establish the pattern of impairment (16). However, these assessments have limitations in the breadth and accuracy of information they are able to provide. Since swallowing is an internal process, ‘bedside’ or clinical assessment do not have the ability to directly observe the structures and physiology involved. Further, some authors have suggested that clinical assessments are insufficient to diagnose aspiration or make adequate recommendations for care in certain populations (17, 18). Therefore, the patient may require an ‘instrumental assessment’.

An instrumental assessment of dysphagia refers to the use of specialist imaging or measurement equipment to investigate the internal mechanisms involved in the swallow. Two are widely considered ‘gold-standards’: the Videofluoroscopic Swallow Study (VFSS) and the Fibreoptic Endoscopy

Evaluation of Swallowing (FEES) (19). The VFSS is the longest-standing instrumental assessment of dysphagia (20). It uses fluoroscopy, a continuous x-ray, to produce a greyscale 'movie' of the oropharynx and oesophagus during the swallowing act. Patients swallow radio-opaque boluses, while the video is recorded for later analysis; a typical VFSS procedure often results in 10 or more individual videos of swallow acts (21). Although developed more recently than the VFSS, the FEES has become a well-established instrumental examination (19). The FEES utilises a flexible nasopharyngo-laryngoscope, passed trans-nasally into the pharynx (22). The patient's swallows are recorded in colour videos and, like the VFSS, an assessment is made of: handling of secretions, food and fluid boluses; the ability to perform swallow manoeuvres; identify the presence of structural abnormalities; and determine the impact of the dysphagia.

This interpretation of recordings produced by VFSS and FEES typically involve the dysphagia clinician viewing the recordings several times and making subjective judgements based on the visuoperceptual features of the images they perceive to be significant. This means that although the FEES and VFSS are frequently referred to as an 'objective' assessment, their interpretation is subjective because there is currently no consensus of standardised criteria to evaluate swallow features (23, 24). One method to overcome this limitation is the use of a measure to interpret video recordings. Measures for FEES and VFFS are typically 'visuoperceptual'. That is, they ascribe ratings to visuoperceptual variables - aspects of the recording which can be interpreted through vision and hearing. These include temporal (perceived duration or timeliness of an event), spatial (perceived location of an event anatomically or scale/size of a clinically relevant indicator), volume (amount of bolus or secretions affected), and patient response variables (such as coughing / choking). In the field of VFSS and FEES, one commonly used example is the Penetration-Aspiration Scale (PAS) (25). This is an eight-point ordinal rating scale which provides descriptors of the penetration and aspiration visualised in VFSS and FEES. Raters select the score they perceive as correlating most closely with patients' performance (e.g., '5: Contrast material contacts the vocal folds but is not ejected').

Although a number of such measures have been reported in the literature, to date there has been no comprehensive systematic review of the FEES and VFSS measures available and their psychometric properties. Comparison across studies, between groups and repeated measures are limited where

measures with questionable psychometric properties are used and diagnosis and decisions about patient care may be compromised.

In a first step to evaluate the quality of the psychometric properties of measures commonly used to analyse VFFS, McCullough et al. (26) reviewed the inter- and intra-rater reliability of the PAS, four measures of duration of swallow events, and nine measures of oropharyngeal function. The authors found that the PAS's intra-rater reliability had better scores than its inter-rater reliability and suggested the inter-reliability of these measures may be unacceptable; they also noted that experienced clinicians had more consistent scores. Frowen et al. (23) examined the psychometric properties of the Bethlehem Assessment Scale (BAS) and ratings of presence / absence of twelve features of swallowing impairments in VFFS. The authors concluded the psychometric properties of these VFSS measures appeared to vary dependent on bolus texture and questioned if the psychometric properties of the VFFS were appropriate for use in clinical and research settings. These studies, while representing a promising start into the investigation of psychometric properties of measures for VFSS, are insufficient to capture the current state of psychometric soundness of VFSS and FEES measures. Further investigation is required.

The COnsensus based Standards for the selection of health Measurement INstruments (COSMIN) checklist (27) provides a taxonomy based in international consensus for the assessment of quality of studies of psychometric properties of measures of aspects of health status or health-related quality of life. Under this taxonomy, methodological quality of studies examining reliability, validity and responsiveness may be examined. To date, this taxonomy has not been applied to studies of measures of VFSS and FEES. The COSMIN has been widely applied to comparable measures; as of June 2014, 560 reviews had been published in PubMed or Embase which had applied the COSMIN to examine measures of health such as delirium, limb function, reflux, spinal injury and sedation (28).

Although the VFSS and FEES are widely considered 'gold-standard' assessments of dysphagia, there are no universally accepted 'gold-standard' measures to interpret them. There is a need for a systematic review of visuoperceptual measures of FEES and VFSS and their psychometric properties based in the COSMIN taxonomy to establish the current state of measures available and lay groundwork for further investigation of their psychometric properties.

Study Aim

There is a lack of comprehensive guidance in the literature regarding measure options for analysis of the FEES and VFSS and their psychometric qualities. Therefore, this study has three aims: 1) to identify visuoperceptual measures which analyse recordings of human swallowing from VFSS and FEES; 2) assess both methodological quality of studies reporting on such measures and the quality of the psychometric properties of these measures and; 3) synthesise this information overall to indicate current state of knowledge about psychometric soundness of visuoperceptual measures of VFSS and FEES. This systematic review focuses on measures that were published in English and assess visuoperceptual aspects of recordings of the VFSS and FEES. It is anticipated that this review will assist in the choice of sound measures to analyse VFSS and FEES by providing an objective account of the psychometric strengths and weaknesses of such measures.

Method

Methodology and reporting of this systematic review was guided by the PRISMA statement. The PRISMA statement is a 27-item checklist required in the transparent reporting of systematic reviews (1). See Supplementary Table 1 for completed PRISMA checklist for the current review.

Eligibility Criteria

Studies eligible for inclusion were research articles which described the psychometric properties of at least one visuoperceptual measure used to analyse VFSS and / or FEES. To be included, studies were required to involve humans any age, visuoperceptual measure/s which analysed data from VFSS or FEES, report on reliability and/or validity of the visuoperceptual measure and be published in English. Studies where measure/s required special software, such as computer programmes which calculate spatial or volume information using pixels, were excluded to better reflect current clinical practices. Although there are several software programmes available to assist recording analysis and offer a more objective interpretation of VFSS and FEES (29), they are limited in terms of clinical use due to the considerable time required to use the software (20). VFSS and FEES clinics typically see

multiple patients consecutively due to limited availability of the equipment and various clinical staff required (30), making routine use of software difficult.

Each instrument was evaluated for reliability and validity according to the COSMIN taxonomy of measurement properties and definitions for health-related patient-reported outcomes (31). However, responsiveness, the ability of a measure to assess change over time, was considered to be outside the scope of this review. = Interpretability, the extent to which qualitative meaning can be ascribed to a measure's quantitative scores or change in scores, was also not considered as this is not regarded as a psychometric property within the COSMIN framework.

Studies which reported only on psychometric properties other than reliability or validity (including responsiveness, interpretability, and/or predictive value), which were published in language other than English, were conference or review papers or unpublished doctoral theses not available online, or where the full scale was unable to be located, were excluded.

Information Sources

A systematic literature search was conducted between 27/01/17 and 10/02/2017 by author RS using four electronic databases: CINAHL, Embase, Medline and Pubmed. Subject headings and free text were used when searching each database, including all dates up until February 2017. Table 1 lists search terms used across all databases. References of articles accepted to the review were hand searched for additional suitable studies.

[Table 1 here]

Study Selection

All abstracts were reviewed by the first author to determine: a) if the study involved human swallowing, b) if an instrumental assessment of swallowing and an associated visuoperceptual measure reporting on the analysis of data arising from the instrumental assessment was present, and c) if the study reported on the psychometric properties of the measure. A random sample of 40% of

abstracts was selected, using an electronic random allocator (www.random.org) and reviewed by a second independent reviewer to establish inter-rater reliability. Abstracts that did not meet two or more of the criteria were excluded from the study. Abstracts which did not meet one of the criteria where discussed by reviewers until consensus was met. Author RS was consulted where consensus could not be reached. Inter-rater reliability was assessed using a quadratic weighting scheme and deemed excellent: Weighted Kappa = 0.895 (95% CI: 0.877 – 0.913). Full texts of acceptable abstracts were retrieved and reviewed. Full texts were likewise excluded if they did not meet criteria (see Figure 2).

Data Collection Process and Data Extraction

Measures fell into two categories: 1) measures with studies which provided information on reliability only, and 2) measures with studies which reported on multiple psychometric properties. Data extracted from studies of measures in the first category were organised under the following descriptive headers: measure, reference, study on psychometrics, aspects evaluated by the measure, summed scores and subscales, total number of items, response options, and the 'domain of variables' assessed by each measure. This final heading was included as it was noted the variables assessed by measures aligned with four broad domains: spatial (e.g., depth of penetration of bolus, range of hyoid movement, spread of secretions), temporal (e.g., time taken for pharyngeal swallow to initiate, time taken to complete oral phase), volume (e.g., amount of residue from boluses, amount of secretions present), and patient response (e.g., no protective airway reflex in response to aspiration).

Measures with studies reporting on more than one psychometric property (e.g., reliability and content validity) also had information extracted under the above categories, with additional data on study purpose and population included, given these studies more comprehensive reporting. Data extracted from these studies was guided by the Cochrane Handbook for Systematic Reviews (32) Section 7.3a and the Systematic Reviews Centre for Reviews and Dissemination (33).

Methodological Quality

The methodological quality of the included studies were assessed using the COSMIN taxonomy of measurement properties and definitions for health-related patient reported outcomes (31, 34). The

COSMIN checklist is a standardised instrument which encompasses nine domains: internal consistency, reliability (including test-retest reliability, inter-rater reliability and intra-rater reliability), measurement error, content validity (including face validity), structural validity, hypotheses testing, cross cultural validity, criterion validity and responsiveness (31). Refer to Table 2 for the definitions of all psychometric properties as defined by the COSMIN statement (34). Criterion validity was not evaluated due to the absence of a 'gold standard' measures for FEES and VFSS. Responsiveness was beyond the scope of this review, and although interpretability is recognised within the COSMIN framework it is not considered a psychometric property and was therefore not assessed. Cross-cultural validity was also not evaluated as all measures reviewed were published in English; where the original measure was developed in a language other than English, quality of translation process was assessed.

[Table 2 here]

Each domain of the COSMIN checklist includes five to 18 items assessing various aspects of study design and statistical analyses. A four-point rating scale designed by Terwee et al. (36) enables an overall methodological quality score to be obtained for each measure, ranging from poor to excellent. Although Terwee et al. (36) recommends making the final quality rating the equivalent of lowest rating of any item in the domain, this makes analysis of subtle differences psychometric qualities of assessments difficult. Therefore a revised scoring system was applied and presented as a percentage: Poor (0-25%), Fair (25.1%-50.0%), Good (50.1%-75%) and Excellent (75.1-100%), as per Cordier et al. (37). As some COSMIN items only have an option to rate as good or excellent, the total score for each psychometric property was calculated using the formula detailed below, to accurately capture the quality of psychometric properties (31):

$$Total\ score\ per\ psychometric\ property = \frac{(Total\ score\ obtained - Min\ score\ possible)}{(Max\ score\ possible - Min\ score\ possible)} \times 100\%$$

After methodological quality of studies was assessed, those which received ratings of 'Excellent', 'Good' and 'Fair' were evaluated using modified criteria by Terwee et al. (36) and Schellingerhout et al. (38), which assesses the quality of the measures' psychometric properties. Studies that received a 'Poor' COSMIN rating were excluded from further analysis, as results arising from studies using doubtful methodology were considered unreliable. Table 3 summarises the criteria used for rating the

quality of content validity, structural validity, hypothesis testing, internal consistency, reliability and measurement error. Finally, each psychometric property for each measure was given an overall score using criteria set out by Schellingerhout (38). An overall quality rating was created by combining the study quality scores measured by COSMIN and the psychometric quality ratings as measured by Terwee et al. (36) and Schellingerhout (38); refer to Table 4. This is consistent with methodology utilised in previous psychometric reviews (39, 40). Refer to Figure 1 flow chart for overview of analysis process.

[Table 3 here]

[Table 4 here]

Data Items, Risk of Bias and Synthesis of Results

Six of the nine COSMIN domains of psychometric properties of each measure were rated from the included publications, with responsiveness and cross-cultural validity excluded. Where an examination of a particular measurement property was not reported in a publication or not described with enough detail to be rated, this was scored as 'not reported' (NR). Risk of bias was addressed with study methodology and psychometric properties of an additional random selection of 40% of studies included in full text being assessed by a second independent reviewer. When scores differed by two points or greater in COSMIN or there was disagreement in Terwee et al. (36) and Schellingerhout et al. (38) ratings, reviewers convened until consensus was achieved. Author RS was consulted to resolve differences in ratings when a consensus could not be reached. Inter-rater reliability for this process was assessed with a weighted Kappa, utilising a quadratic scheme. Results indicated excellent agreement (Weighted Kappa: 0.897, 95% CI: 0.867-0.927). Tables 5, 6 and 7 displays the synthesised data collected from each measure and article reporting on psychometric properties.

[Figure 1 here]

Results

Systematic Literature Search

A total of 2,090 abstracts were retrieved from database searches, including duplicates. Abstracts per database were: CINAHL = 108, Embase = 298, Medline = 255, PubMed = 1,429. Abstract duplicates totalled 293. Duplicates were removed and 1,797 abstracts were screened for inclusion in the review, with 1,581 being rejected. Subsequently 216 full text articles were assessed for eligibility. Reference lists of included studies were also searched for additional studies. Of these, 45 studies encompassing 39 measures met the inclusion criteria. Figure 2 illustrates the reviewing process according to PRISMA and details abstract and full text exclusion reasons.

[Figure 2 here]

Included Measures

Due to the limited information available about their psychometric properties, measures where information is available solely on reliability are presented separately (Table 5) from the measures with information about multiple psychometric properties (Tables 6 and 7). These were collated separately, as measures with known psychometric properties for both reliability and validity are likely to be more relevant to the clinician or researcher. Table 5 synthesises the characteristics of these 21 reliability-only measures. Six measures analysed FEES recordings only; 14 measures were for VFSS recordings and one analysed both FEES and VFSS recordings (i.e., 7 measures of FEES and 15 measures of VFSS). FEES measures most commonly included the variables related to aspiration, penetration, secretions and residue (5 of 7), while VFSS measures most commonly had variables related to pharyngeal residue (10 of 15), aspiration (8 of 15), timing of swallow initiation (7 of 15) pharyngeal phase duration (7 of 15) and oral phase duration (6 of 15). Oesophageal parameters (such as reflux, bolus stasis, Zenker's diverticulum) were the most uncommon variables, with only two of the 15 measures reporting on oesophageal characteristics. None of the measures utilised summed scores or subscales; all were comprised of one or more single variables. With the exception of Gosa et al. (41), all studies recruited adult populations only. Overall, the majority of measures (16 of 21) were created by the authors of the same study which reported on their psychometrics. Measures were considered to have been created by the authors when: 1) authors reported selecting the measure's variables from the literature without reference to an earlier measure utilising these variables, and/or 2) authors indicated the measure was created at their facility or for the purposes of their study.

Across both FEES and VFSS measures, the most commonly used response options were nominal scales (*n* = 10) and ordinal scales with associated descriptors at each level (*n* = 9; e.g., secretion colour: clear, white, brown, yellow or bloody' and '0 = no pooling, 1 = filling of <50% of the vallculae, 2 = filling of >50% of valleculae). Other options included dichotomous scales (*n* = 6; e.g., aspiration present: yes / no), and open-ended response options, where raters recorded their judgements of continuous variables, such as time taken to complete a swallow phase (*n* = 6). The number of items utilised in FEES measures ranged from one to 16 (mean = 4.4). VFSS measures used a greater range, from one to 23 (mean = 8.3). Overall, 16 measures used less items than the mean for their respective instrumental assessment; of these, eight scored overall positive for reliability (42-49), five had conflicting results (50-54), two negative (43, 44) and one indeterminate (55). Six measures used more items than the mean; none scored positive for reliability overall. Two of the six received conflicting ratings (26, 56) and two negative (57, 58), one scored 'indeterminate' (41), and one study was not evaluated due to 'poor methodological quality' (59). It should also be noted that two studies reported reliability for two different protocols (green coloured boluses vs. white) and diagnoses (aspiration or dysphagia) (43, 44); both scored positive for reliability overall in only one protocol or diagnosis (green bolus and dysphagia respectively).

Table 6 describes the characteristics of the 18 measures with known multiple psychometric properties or properties other than reliability only. Seven measures analysed FEES recordings only and eight measures analysed VFSS recordings only; three measures pertained to both FEES and VFSS. This resulted in 10 measures for FEES and 11 measures for VFSS.

FEES measures most commonly evaluated amount or colour of secretions / residue (*n* = 10). Two measures assessed penetration / aspiration, with patient response to airway invasion assessed by three measures. Two measures utilised a summed score or subscales to formulate overall ratings: P-Score (60) and the BRACS (61). The remainder did not use summed scores / subscales. Among measures of VFSS the most commonly analysed variables were pharyngeal residue (*n* = 9), swallow reflex initiation (*n* = 5), penetration / aspiration (*n* = 4), oral transit duration (*n* = 5), laryngeal / hyoid elevation (*n* = 4), pharyngeal transit duration (*n* = 4), bolus formation / control (*n*= 4), epiglottic movement (*n* = 4), and lip closure (*n* = 3). Similar to measures that reported on reliability only (Table

5), function of oesophagus was the most rarely included variable in the assessment, with only one measure including analysis of the oesophageal phase swallow (62). Consistent with FEES measures, VFSS measures also rarely utilised subscales or summed scores. A total of three measures included summed overall scores [FDS (63), VDS (64), Unnamed - Single variable - Residue, (65)], while two utilised subscales [MBSImp (62) and DIGEST (66)].

Among measures of FEES, total number of items ranged from one to 16 (mean = 3.7). The number of items utilised in VFSS measures was slightly higher, ranging from one to 17 (mean = 6.5). Response options in FEES measures were most commonly ordinal (*n* = 8) and ranged from 3- to 8-point scales. Two measures used nominal response scales. Conversely, nominal scales, were more common among VFSS measures (*n* = 6). They used a range of criterion such as volume / severity descriptors (e.g., 'absent, trace / minimal, moderate / maximal, unable to visualise' or 'none, <10%, 10- 50%, >50%'). Ordinal scales (*n* = 4) ranging from 2- to 8-points, dichotomous scales (*n* = 3), and continuous response options such as time (*n* = 2) were used less frequently in VFSS. Two measures used multiple types of response options (67, 68).

Table 7 synthesises information from the 29 studies which examined the 18 measures with multiple psychometric data. The majority of measures had their psychometrics investigated by only one study (*n* = 13). All but one study examined adult populations; one included children and adults (69). Age varied widely, from 10 – 100 years (mean = 61.4 years; SD = 7.7). Aetiology similarly varied widely and included acquired neurological conditions, neurodegenerative diseases, head and neck cancers, pulmonary and cardiac conditions and trauma (acquired brain injury, burns, non-specific traumas). The most common diagnostic groups included by studies were stroke (*n* = 25), degenerative neurological diseases (*n* = 14) and head and neck cancers (*n* = 10). Number of participants studied ranged from 13 to 1,995 (mean = 161.6 [SD = 376.7]; median = 45 [IQR = 80]). According to the COSMIN taxonomy, recruitment of more than 100 participants are recommended to explore internal consistency, reliability, measurement error and hypothesis testing. The median number of participants included in the data set indicates most studies used sample sizes that were less than ideal. Where validation studies use a limited sample size, the accuracy of their conclusions and generalisability of results to the wider population is questionable.

[Table 5 here]

[Table 6 here]

[Table 7 here]

Psychometric Properties

Table 5 summarises the quality ratings of 21 measures where information is available about reliability only. According to COSMIN ratings, one study had 'Poor' methodological quality (which was excluded from further analysis), nine 'Fair', 10 'Good' and one 'Excellent'. The overall quality ratings, based on Terwee et al. (36) and Schellingerhout et al. (38), resulted in two measures with moderate negative ratings, two with limited negative, two indeterminate, three with limited positive evidence, six with moderate positive scores and seven with conflicting ratings.

The methodology quality ratings of studies (as determined by COSMIN), which report on more than one psychometric property or properties other than reliability only, are described in Table 8. Included articles most commonly reported on reliability (*n* = 22) and hypothesis testing (*n* = 17). In addition, one study reported on internal consistency, 12 on content validity and two on structural validity. No studies described measurement error. Measures which utilised only one item could not be assessed for internal consistency; this property is marked not applicable (N/A) for these studies. Although all studies were published in English, it is likely two measures were developed in another language (74, 78). Authors were contacted to clarify the translation process and quality of the translation process to English was assessed, using the COSMIN ratings of cross cultural validity. Table footnotes provide further description of these measures. The ratings of the quality of studies of measures varied considerably across psychometric properties. Study quality for structural validity ranged from good to excellent, while content validity, internal consistency and reliability ranged from poor to excellent. Hypothesis testing results ranged from poor to fair. Properties of measures which received a poor rating (*n* = 3) were excluded from further analysis.

Table 9 provides a summary of the quality of psychometric properties of included measures based on Terwee et al. (36) and Schellingerhout et al. (38), whereas Table 10 summarises of the overall quality

ratings per psychometric property of nine FEES measures and nine VFSS measures, as evaluated against Schellingerhout et al (38) criteria. One measure, PAS (25), assessed both FEES and VFSS; as such, the results were reported separately as it had different psychometric properties for FEES and VFSS respectively. The notes section of Table 10 provides a description of the criteria used to rate the overall psychometric quality. Reliability was the most commonly (*n* = 14) assessed psychometric property, followed by hypothesis testing (*n* = 13) and content validity (*n* = 12). Structural validity was analysed twice and one study reported on internal consistency. Each measure had between two and four psychometric properties present. Only eight measures were found to have one or more properties with positive psychometric soundness (60, 61, 63, 64, 66, 69, 74, 75). Four measures had conflicting evidence (21, 25, 66). One measure had limited negative evidence (64). The most frequent finding was indeterminate (*n* = 27). Overall, information about psychometric properties was very limited, with no measures emerging as strong over a range of properties.

[Table 8 here]

[Table 9 here]

[Table 10 here]

Discussion

The purpose of this review was to identify visuoperceptual measures for analysing the 'gold-standard' instrumental assessments of dysphagia, FEES and VFSS, and to evaluate the psychometric robustness of these measures. Comprehensive assessment of dysphagia often involves instrumental assessment; however, the data which are produced through these assessments are not meaningful in and of itself. It must be interpreted by the dysphagia clinician in a manner which is accurate, consistent, and appropriate to purpose to guide diagnosis and management. This systematic review identified 39 visuoperceptual measures from 45 research articles that are used by researchers and practitioners to interpret the FEES and VFSS recordings. The COSMIN checklist, which appraises the quality of the studies, was used in combination with quality criteria of the psychometric properties as described by Terwee et al. (36) and Schellingerhout et al. (38). Evaluation using the COSMIN

taxonomy enabled a standardised and thorough approach to examination of the quality of psychometrics of these measures (27, 88). This systematic review therefore provides a comprehensive summary of the quality of psychometric properties of visuoperceptual measures currently available for VFSS and FEES.

Psychometric quality of measures overall

A total of 18 measures reported on more than one psychometric property or properties other than reliability only, while 21 measures reported solely on reliability. Data about the psychometric properties of the 18 measures were found on internal consistency, reliability, content validity, structural validity and hypothesis testing. Information was most frequently available on reliability (intra and inter-rater), content validity and hypothesis testing; only two measures reported data on structural validity (61, 62), and one on internal consistency (61). Where information is lacking on internal consistency and structural validity, it cannot be assumed the items within the measure are all manifestations of the underlying construct and that the scores of the measure reflect the dimensionality of the construct. For example, a measure for VFSS which has a number of items, arbitrarily evenly separated into subscales of oral, pharyngeal and oesophageal phases, may have items placed in the incorrect categories. Therefore, a clinician may be scoring items which are ostensibly placed in the oesophageal phase, but which in fact represent pharyngeal phase dysfunction. This may change diagnosis and management approach (e.g., unnecessary referral onwards to gastroenterology). No studies reported on the property 'measurement error'. Measurement error assess whether changes in scores are related to true change in the construct of interest or other random factors. Inadequate information on this property means it cannot be assumed that alteration in a patient's scores indicate improving or worsening dysphagia versus changes other related factors.

The most common overall result across all of the assessed psychometric properties was 'indeterminate' (64%). 'Indeterminate' indicates neither positive nor negative findings; it is a marker that further information or research is required. 'Indeterminate' ratings were particularly common in hypothesis testing; all 13 measures that reported on hypothesis testing received 'indeterminate' ratings. Hypothesis testing examines the relationship of the measure compared to other measures, or difference between groups. Specific hypotheses should be formulated a-priori, with expected direction

and magnitude of correlations stated (88). An example would be: 'We expect *x-measure* of residue to correlate positively with *y-measure* of residue, (r >0.70).' None of the studies clearly formulated their hypotheses a-priori and stated expected direction and magnitude of correlations. This issue with reporting and research formulation resulted in the high rates of 'indeterminate' overall scores.

Content validity was another psychometric property with high rates of 'indeterminate' findings. Content validity is the relevance and comprehensiveness of items within a measure. To establish adequate content validity, it is recommended that experts should judge the relevance of the items. Comprehensiveness of items should be established by providing a clear theoretical foundation for the item selection. Assessment should also be completed of whether all relevant aspects of a construct are subsumed within the measure (88). The content validity ratings of measures included in this review was negatively affected by lack of reference to expert groups (e.g., lack of use of the Delphi technique to establish expert consensus), lack of clear description of the experts involved in the formulation of the measure, lack of clear description of the target population and concepts that are being measures and, in some cases, the absence of any reference to literature to explain the selection of items used in the conceptualisation of the measure. Deficiencies in establishing and reporting on content validity has significant clinical implications; it is unclear what such measures are in fact measuring. The measure may be unfit for particular clinical purposes or populations, or the entire measures may be problematic and unsuitable for use. In addition to common 'indeterminate' results, 'limited' strength of evidence was also a frequent finding (17%). This was the result of the low rate of psychometric properties investigated per study for each measure and most measures (31 of the 39 measures), conducted only one study to investigate a single psychometric property. This suggests more research of adequate design and methodological quality is required to report on these psychometric properties.

Measure design and characteristics

Predominantly, measures of VFSS examined pharyngeal residue, penetration / aspiration, timing of pharyngeal initiation, oral and pharyngeal phase duration and laryngeal / hyoid elevation. FEES measures most commonly reported on, residue penetration / aspiration and secretions. This is likely a

reflection of seminal works on the use and analysis of the FEES and VFSS (73, 89) and the importance of aspiration as a predictor of aspiration pneumonia and chronic dysphagia (90, 91).

None of the studies described how response options were designed or decisions on the number of items was made. Measure design may have had an impact on the quality of psychometric properties; analysis of overall scores of measures with reliability data only revealed use of fewer items appeared to correspond with increased reliability scores. It was also noted VFFS measures on average used three more items than FEES measures and the upper range of items used was higher (23 versus 16 respectively). VFSS measures generally used nominal scales, while FEES measures used ordinal scales. Of note, VFSS measures scored less positively overall compared with FEES measures; the greater complexity of response options and number of items may have affected in this outcome.

Among the 18 measures which reported on psychometric properties other than solely reliability (Table 6 – 10), only seven utilised subscales and / or summed scores (60-66). Use of composite scores allows examinations of dimensions (inter-related variables) and comparison between constructs; measures which do not use subscales or summed scores may be less comprehensive than those that do. Across all studies included in this review, only two utilised paediatric populations (41, 69). This highlights an urgent need for studies which explore of the psychometrics of visuoperceptual measures of FEES and VFSS that are used in paediatric populations.

Theoretical models

Classical Test Theory (CTT) was the underlying theoretical model used in all studies included in this review; none of studies used Item Response Theory (IRT). CTT makes assumptions of item equivalence and of standard error of measurement (92). These assumptions may impact ordinal and nominal scales; for example, the assumption that a grade of 3 in a 5-point scale is an exact mid-point of severity may be inaccurate. Grades within scales may in fact carry different weights. In addition, a significant limitation of CTT is its relatively weak theoretical assumptions and circular dependency, specifically: a) the person statistic (i.e., observed score) is item sample dependent; and b) the item statistics are examinee/person sample dependent, which poses some difficulties in CTT's application in some measurement situations (93). IRT was developed in response to some of the limitations of

CTT. IRT also has limitations; it is a complex model which requires much larger samples of participants and items compared to CTT (94). Although the COSMIN taxonomy does not specify superiority of either model, IRT methods are increasingly being utilised for the development of assessments within fields such as psychology and have numerous reported advantages over CTT only methods (95, 96). It is beyond the scope of this review to conduct an in-depth discussion of the theoretical statistical frameworks utilised by measures in this study; however, it is suggested further investigation is needed to examine reasons for the lack of IRT methods in measures of VFSS and FEES and relative strengths and appropriateness of the models to this field.

Psychometric properties of measures with relative strength of evidence

The available information on all measure's psychometric properties was extremely limited. Therefore, although some measures appear to have stronger evidence in relation to others, this is based on a very small data pool. Of the measures where data were available, the measures for FEES which scored the strongest levels of evidence overall were the BRACS (61) and the Dysphagia Score (74); BRACS scored moderately positive for reliability and structural validity, while the Dysphagia score had limited positive evidence of reliability and content validity. As information about only two measurement properties were available, information on measure quality, while indicating relative strength, should be considered incomplete. The BRACS received scores of indeterminate for internal consistency, content validity and hypothesis testing categories due to a small sample size, unclear description of item and concept selection, and lack of a-priori hypotheses respectively. The measure would benefit from further research utilising a larger sample size (> 100) and addressing these reporting issues. Measurement error should also be investigated. The Dysphagia Score would benefit from further research investigating intra-rater reliability, more detailed reporting of how construct validity was ensured and assessment to determine if all items are relevant to the constructs being measured. Properties of internal consistency, measurement error, structural validity, and hypothesis testing should be investigated in future research.

In terms of VFSS analysis, the DIGEST (66) had the highest rated evidence overall, with strongly positive content validity. An indeterminate score was recorded in hypothesis testing due to lack of a-priori hypotheses, and conflicting reliability was found due to positive intra-rater reliability but negative

intra-rater reliability (weighted K <0.70). The DIGEST would benefit from further research investigating its psychometrics, specifically internal consistency, measurement error, and structural validity. As with the FEES measures, although the DIGEST exhibits relative strength of evidence, there are significant gaps in data on its psychometrics and its ranking as a 'stronger' measure has noteworthy caveats.

No other measures with multiple known psychometrics in VFSS had moderate levels of evidence. Of the measures with reliability data known only, the BAS (70), an unnamed 'presence / absence of aspiration' dichotomous scale (42), an unnamed scale of temporal and spatial variables (45), and an unnamed scale of temporal variables (46) had moderate positive evidence of reliability. However, positive findings in reliability do not mean the measure has appropriate validity; further assessment of these measures is required.

Overall, even though some measures of FESS and VFSS recordings had higher levels of evidence of psychometric quality compared with other measures, the findings are based on very limited information about psychometric qualities and limited numbers of studies on psychometric properties. This lack of data is striking, given the ubiquitous use of instrumental assessment in dysphagia research and clinical management. Overall, significantly more research is needed on the psychometric properties of measures.

Limitations

Although every effort was taken to ensure the scientific rigour of this systematic review, there were a number of limitations that should be acknowledged. It should be noted the authors of this review did not contact authors of the studies included in this review for missing data; consequently, some information may not have been included. Further, evaluating the qualities of criterion validity and responsiveness was not attempted in this review. Criterion validity was not attempted as there is no acknowledged gold-standard measure to use as a benchmark. Inclusion of responsiveness would have necessitated analysis of all studies which utilise visuoperceptual outcome measures, which would have made the size of this review unmanageable. However, it is acknowledged responsiveness is an important psychometric property which would benefit from detailed review in the future.

Conclusion

Accurate assessment and diagnosis of the pathology of swallowing impairments using instrumental assessments is an important part of practice for most clinicians and researcher working within the field of dysphagia. Therefore, it is important that the measures which analyse the data these instruments generate are psychometrically sound. This review assessed the reliability and validity of visuoperceptual measures for FEES and VFSS. In the context of significant gaps and in the evidence regarding psychometric quality for all measures, it was concluded the BRACS, Dysphagia score and the DIGEST had indications of adequate evidence for some psychometrics properties. Notably, even though these measures show relative promise, their psychometric quality and the quality of all measures retrieved overall was relatively weak. In addition, no measure had complete information about all of its psychometric properties available. This is likely related to the lack of studies on the psychometrics of measures and the narrow range of properties investigated within these studies. Most measures were examined in one study only, which did not comprehensively assess all psychometric properties.

The findings from this systematic review has direct clinical implications; these measures represent the options available for clinical practice, however very little is known about their properties. This means their validity and suitability for use in practice and research settings may be limited and questionable. Overall, there is insufficient evidence to recommend any individual measure included in this review as valid and reliable to interpret VFSS and FEES generated recordings. Further research is required to investigate the psychometric properties of the measures that have not been evaluated to date. This review highlights the need for studies reporting on the psychometrics of visuoperceptual measures for FEES and VFSS which utilise more robust psychometric methodological designs, including using adequate sample sizes and appropriate statistical analyses, and which adopts appropriate study designs and reporting practices.

Supporting Information

S1 Table. PRISMA checklist for the current systematic review. From Moher D, Liberati A, Tetzlaff J,
Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097.

Author Contributions

[blinded for review]

References

1. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of Internal Medicine. 2009;151(4):264-9.
2. Mercadante S, Aielli F, Adile C, Ferrera P, Valle A, Fusco F, et al. Prevalence of oral mucositis, dry mouth, and dysphagia in advanced cancer patients. Supportive Care in Cancer. 2015;23(11):3249-55.
3. Takizawa C, Gemmell E, Kenworthy J, Speyer R. A systematic review of the prevalence of oropharyngeal dysphagia in stroke, Parkinson's disease, Alzheimer's disease, head injury, and pneumonia. Dysphagia. 2016;31(3):434-41.
4. Dodrill P, Gosa MM. Pediatric dysphagia: physiology, assessment, and management. Annals of Nutrition and Metabolism. 2015;66(Suppl. 5):24-31.
5. Kalf J, De Swart B, Bloem B, Munneke M. Prevalence of oropharyngeal dysphagia in Parkinson's disease: a meta-analysis. Parkinsonism & related disorders. 2012;18(4):311-5.
6. Kertscher B, Speyer R, Fong E, Georgiou AM, Smith M. Prevalence of oropharyngeal dysphagia in the Netherlands: a telephone survey. Dysphagia. 2015;30(2):114-20.
7. Holland G, Jayasekeran V, Pendleton N, Horan M, Jones M, Hamdy S. Prevalence and symptom profiling of oropharyngeal dysphagia in a community dwelling of an elderly population: a self-reporting questionnaire survey. Diseases of the Esophagus. 2011;24(7):476-80.
8. Park Y-H, Han H-R, Oh B-M, Lee J, Park J-a, Yu SJ, et al. Prevalence and associated factors of dysphagia in nursing home residents. Geriatric Nursing. 2013;34(3):212-7.
9. Cichero JA, Heaton S, Bassett L. Triaging dysphagia: nurse screening for dysphagia in an acute hospital. Journal of clinical nursing. 2009;18(11):1649-59.
10. Yi S-H, Kim S-J, Huh J, Jun T-G, Cheon HJ, Kwon J-Y. Dysphagia in infants after open heart procedures. American journal of physical medicine & rehabilitation. 2013;92(6):496-503.
11. DiBardino DM, Wunderink RG. Aspiration pneumonia: a review of modern trends. Journal of critical care. 2015;30(1):40-8.
12. Komiya K, Ishii H, Umeki K, Mizunoe S, Okada F, Johkoh T, et al. Impact of aspiration pneumonia in patients with community-acquired pneumonia and healthcare-associated pneumonia: A multicenter retrospective cohort study. Respirology. 2013;18(3):514-21.

13. Garcia-Peris P, Parón L, Velasco C, De la Cuerda C, Camblor M, Bretón I, et al. Long-term prevalence of oropharyngeal dysphagia in head and neck cancer patients: impact on quality of life. Clinical Nutrition. 2007;26(6):710-7.
14. Leow LP, Huckabee M-L, Anderson T, Beckert L. The impact of dysphagia on quality of life in ageing and Parkinson's disease as measured by the swallowing quality of life (SWAL-QOL) questionnaire. Dysphagia. 2010;25(3):216-20.
15. Verdonschot RJ, Baijens LW, Serroyen JL, Leue C, Kremer B. Symptoms of anxiety and depression assessed with the Hospital Anxiety and Depression Scale in patients with oropharyngeal dysphagia. Journal of psychosomatic research. 2013;75(5):451-5.
16. Luker JA, Wall K, Bernhardt J, Edwards I, Grimmer-Somers K. Measuring the quality of dysphagia management practices following stroke: a systematic review. International Journal of Stroke. 2010;5(6):466-76.
17. McCullough G, Rosenbek J, Wertz R, McCoy S, Mann G, McCullough K. Utility of clinical swallowing examination measures for detecting aspiration post-stroke. Journal of Speech, Language, and Hearing Research. 2005;48(6):1280-93.
18. Carnaby-Mann G, Lenius K. The bedside examination in dysphagia. Physical Medicine and Rehabilitation Clinics of North America. 2008;19(4):747-68.
19. Langmore SE. History of Fiberoptic Endoscopic Evaluation of Swallowing for Evaluation and Management of Pharyngeal Dysphagia: Changes over the Years. Dysphagia. 2017:1-12.
20. Huckabee M-L, Macrae P, Lamvik K. Expanding instrumental options for dysphagia diagnosis and research: ultrasound and manometry. Folia Phoniatrica et Logopaedica. 2015;67(6):269-84.
21. Karnell MP, Rogus NM. Comparison of Clinician Judgments and Measurements of Swallow Response TimeA Preliminary Report. Journal of Speech, Language, and Hearing Research. 2005;48(6):1269-79.
22. Dziewas R, Glahn J, Helfer C, Ickenstein G, Keller J, Ledl C, et al. Flexible endoscopic evaluation of swallowing (FEES) for neurogenic dysphagia: training curriculum of the German Society of Neurology and the German stroke society. BMC Medical Education. 2016;16(1):70.
23. Frowen JJ, Cotton SM, Perry AR. The stability, reliability, and validity of videofluoroscopy measures for patients with head and neck cancer. Dysphagia. 2008;23(4):348-63.

24. Rommel N, Hamdy S. Oropharyngeal dysphagia: manifestations and diagnosis. Nature reviews Gastroenterology & Hepatology. 2016;13(1):49.

25. Rosenbek JC, Robbins JA, Roecker EB, Coyle JL, Wood JL. A penetration-aspiration scale. Dysphagia. 1996;11(2):93-8.

26. McCullough GH, Wertz RT, Rosenbek JC, Mills RH, Webb WG, Ross KB. Inter-and intrajudge reliability for videofluoroscopic swallowing evaluation measures. Dysphagia. 2001;16(2):110-8.

27. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010;19(4):539-49.

28. Terwee CB. An overview of systematic reviews of measurement properties of outcome measurement instruments that intend to measure (aspects of) health status or (health- related) quality of life. Department of Epidemiology and Biostatistics VU University Medical Center Amsterdam, the Netherlands: The COSMIN group, 2014 2014. Report No.

29. Pearson WG, Molfenter SM, Smith ZM, Steele CM. Image-based measurement of post-swallow residue: the normalized residue ratio scale. Dysphagia. 2013;28(2):167-77.

30. Newman RD, Nightingale J. Improving patient access to videofluoroscopy services: Role of the practitioner-led clinic. Radiography. 2011;17(4):280-3.

31. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology. 2010;10(22):1-8.

32. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews for Interventions.: Wiley Online Library; 2008.

33. Centre for Reviews Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. Layerthorpe, York.: CRD University of York; 2009.

34. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. International consensus on taxonomy, terminology and definitions of measurement properties for health related patient reported outcomes: results of the COSMIN study. Journal of Clinical Epidemology. 2010;63:737-45.

35. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010;63(7):737-45.

36. Terwee CB, Bot S, de Boer M, van der Windt D, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionaires. Journal of Clinical Epidemiology. 2007;60:34-42.

37. Cordier R, Speyer R, Chen YW, Wilkes-Gillan S, Brown T, Bourke-Taylor H. Evaluating the psychometric quality of social skills measures: A systematic review. Plos One. 2015;10(7).

38. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet H, Terwee CB. Measurement properties of disease-specific questionaires in patients with neck pain: a systematic review. Quality of Life Research. 2012;21:659-70.

39. Author. (2008). [Title omitted for blind review]. Plos One. 2016;11(1):1-24.

40. Author. (2008). [Title omitted for blind review]. Plos One. 2016.

41. Gosa MM, Suiter DM, Kahane JC. Reliability for identification of a select set of temporal and physiologic features of infant swallows. Dysphagia. 2015;30(3):365-72.

42. Hind JA, Gensler G, Brandt DK, Gardner PJM, Blumenthal L, Gramigna GD, et al. Comparison of trained clinician ratings with expert ratings of aspiration on videofluoroscopic images from a randomized clinical trial. Dysphagia. 2009;24(2):211.

43. Mann G, Hankey GJ, Cameron D. Swallowing disorders following acute stroke: prevalence and diagnostic accuracy. Cerebrovascular diseases. 2000;10(5):380-6.

44. Marvin S, Gustafson S, Thibeault S. Detecting aspiration and penetration using FEES with and without food dye. Dysphagia. 2016;31(4):498-504.

45. Nordin NA, Miles A, Allen J. Measuring Competency Development in Objective Evaluation of Videofluoroscopic Swallowing Studies. Dysphagia. 2017;32(3):427-36.

46. Power ML, Hamdy S, Goulermas JY, Tyrrell PJ, Turnbull I, Thompson DG. Predicting aspiration after hemispheric stroke from timing measures of oropharyngeal bolus flow and laryngeal closure. Dysphagia. 2009;24(3):257-64.

47. Rommel N, Borgers C, Van Beckevoort D, Goeleven A, Dejaeger E, Omari TI. Bolus Residue Scale: An easy-to-use and reliable videofluoroscopic analysis tool to score bolus residue in patients with dysphagia. International Journal of Otolaryngology. 2015;2015.

48. Susa C, Kagaya H, Saitoh E, Baba M, Kanamori D, Mikushi S, et al. Classification of Sequential Swallowing Types Using Videoendoscopy with High Reproducibility and Reliability. American Journal of Physical Medicine & Rehabilitation. 2015;94(1):38-43.

49. Warnecke T, Suttrup I, Schröder JB, Osada N, Oelenberg S, Hamacher C, et al. Levodopa responsiveness of dysphagia in advanced Parkinson's disease and reliability testing of the FEES-Levodopa-test. Parkinsonism & Related Disorders. 2016;28:100-6.

50. Pilz W, Vanbelle S, Kremer B, van Hooren MR, van Becelaere T, Roodenburg N, et al. Observers' agreement on measurements in fiberoptic endoscopic evaluation of swallowing. Dysphagia. 2016;31(2):180-7.

51. Kelly A, Leslie P, Beale T, Payten C, Drinnan M. Fibreoptic endoscopic evaluation of swallowing and videofluoroscopy: does examination type influence perception of pharyngeal residue severity? Clinical Otolaryngology. 2006;31(5):425-32.

52. Gibson E, Phyland D, Marschner I. Rater reliability of the modified barium swallow. Australian Journal of Human Communication Disorders. 1995;23(2):54-60.

53. Lee JW, Randall DR, Evangelista LM, Kuhn MA, Belafsky PC. Subjective Assessment of Videofluoroscopic Swallow Studies. Otolaryngology–Head and Neck Surgery. 2017;156(5):901-5.

54. Miles A. Inter-rater reliability for speech–language therapists' judgement of oesophageal abnormality during oesophageal visualization. International Journal of Language & Communication Disorders. 2016.

55. Rodriguez KH, Roth CR, Rees CJ, Belafsky PC. Reliability of the pharyngeal squeeze maneuver. Annals of Otology, Rhinology & Laryngology. 2007;116(6):399-401.

56. Tohara H, Nakane A, Murata S, Mikushi S, Ouchi Y, Wakasugi Y, et al. Inter-and intra-rater reliability in fibroptic endoscopic evaluation of swallowing. Journal of oral rehabilitation. 2010;37(12):884-91.

57. Bryant KN, Finnegan E, Berbaum K. VFS interjudge reliability using a free and directed search. Dysphagia. 2012;27(1):53-63.

58. Stoeckli SJ, Huisman TA, Seifert BA, Martin–Harris BJ. Interrater reliability of videofluoroscopic swallow evaluation. Dysphagia. 2003;18(1):53-7.

59. Scott A, Perry A, Bench J. A study of interrater reliability when using videofluoroscopy as an assessment of swallowing. Dysphagia. 1998;13(4):223-7.

60. Farneti D. Pooling score: an endoscopic model for evaluating severity of dysphagia. Acta Otorhinolaryngologica Italica. 2008;28(3):135.

61. Kaneoka AS, Langmore SE, Krisciunas GP, Field K, Scheel R, McNally E, et al. The Boston residue and clearance scale: preliminary reliability and validity testing. Folia Phoniatrica et Logopaedica. 2013;65(6):312-7.

62. Martin-Harris B, Brodsky MB, Michel Y, Castell DO, Schleicher M, Sandidge J, et al. MBS measurement tool for swallow impairment—MBSImp: establishing a standard. Dysphagia. 2008;23(4):392-405.

63. Han TR, Paik N-J, Park JW. Quantifying swallowing function after stroke: a functional dysphagia scale based on videofluoroscopic studies. Archives of physical medicine and rehabilitation. 2001;82(5):677-82.

64. Han TR, Paik N-J, Park J-W, Kwon BS. The prediction of persistent dysphagia beyond six months after stroke. Dysphagia. 2008;23(1):59-64.

65. Omari TI, Dejaeger E, Van Beckevoort D, Goeleven A, De Cock P, Hoffman I, et al. A novel method for the nonradiological assessment of ineffective swallowing. The American Journal of Gastroenterology. 2011;106(10):1796-802.

66. Hutcheson KA, Barrow MP, Barringer DA, Knott JK, Lin HY, Weber RS, et al. Dynamic Imaging Grade of Swallowing Toxicity (DIGEST): scale development and validation. Cancer. 2017;123(1):62-70.

67. Daniels SK, Schroeder MF, McClain M, Corey DM. Dysphagia in stroke: development of a standard method to examine swallowing recovery. Journal of rehabilitation research and development. 2006;43(3):347.

68. Karnell MP, Rogus NM. Comparison of clinician judgments and measurements of swallow response time: A preliminary report. Journal of Speech, Language, and Hearing Research. 2005;48(6):1269-79.

69. Donzelli J, Wesling M, Brady S, Craney M. Predictive value of accumulated oropharyngeal secretions for aspiration during video nasal endoscopic evaluation of the swallow. Annals of Otology, Rhinology & Laryngology. 2003;112(5):469-75.
70. Scott AG. The development of a scale to assess swallowing function in motor neuron disease using videofluoroscopic techniques: La Trobe University; 1999.
71. Leonard RJ, Kendall KA, McKenzie S, Gonçalves MI, Walker A. Structural displacements in normal swallowing: a videofluoroscopic study. Dysphagia. 2000;15(3):146-52.
72. McCullough GH, Wertz RT, Rosenbek JC, Dinneen C. Clinicians' preferences and practices in conducting clinical/bedside and videofluoroscopic swallowing examinations in an adult, neurogenic population. American Journal of Speech-Language Pathology. 1999;8(2):149-63.
73. Leonard R, Kendall K. Dysphagia assessment and treatment planning: a team approach: Cengage Learning; 1997.
74. Dziewas R, Warnecke T, Ölenberg S, Teismann I, Zimmermann J, Krämer C, et al. Towards a basic endoscopic assessment of swallowing in acute stroke–development and evaluation of a simple dysphagia score. Cerebrovascular Diseases. 2008;26(1):41-7.
75. Neubauer PD, Rademaker AW, Leder SB. The Yale Pharyngeal Residue Severity Rating Scale: an anatomically defined and image-based tool. Dysphagia. 2015;30(5):521-8.
76. Murray J, Langmore SE, Ginsberg S, Dostie A. The significance of accumulated oropharyngeal secretions and swallowing frequency in predicting aspiration. Dysphagia. 1996;11(2):99-103.
77. Curtis JA, Laus J, Yung KC, Courey MS. Static endoscopic evaluation of swallowing: transoral endoscopy during clinical swallow evaluations. The Laryngoscope. 2016;126(10):2291-4.
78. Park WY, Lee TH, Ham NS, Park JW, Lee YG, Cho SJ, et al. Adding endoscopist-directed flexible endoscopic evaluation of swallowing to the videofluoroscopic swallowing study increased the detection rates of penetration, aspiration, and pharyngeal residue. Gut and liver. 2015;9(5):623.
79. Farneti D, Fattori B, Nacci A, Mancini V, Simonelli M, Ruoppolo G, et al. The Pooling-score (P-score): inter-and intra-rater reliability in endoscopic assessment of the severity of dysphagia. ACTA otorhinolaryngologica italica. 2014;34(2):105.
80. Pluschinski P, Zaretsky E, Stöver T, Murray J, Sader R, Hey C. Validation of the secretion severity rating scale. European Archives of Oto-Rhino-Laryngology. 2016;273(10):3215-8.

81. Butler SG, Stuart A, Case LD, Rees C, Vitolins M, Kritchevsky SB. Effects of liquid type, delivery method, and bolus volume on penetration-aspiration scores in healthy older adults during flexible endoscopic evaluation of swallowing. Annals of Otology, Rhinology & Laryngology. 2011;120(5):288-95.

82. Butler SG, Markley L, Sanders B, Stuart A. Reliability of the penetration aspiration scale with flexible endoscopic evaluation of swallowing. Annals of Otology, Rhinology & Laryngology. 2015;124(6):480-3.

83. Colodny N. Interjudge and intrajudge reliabilities in fiberoptic endoscopic evaluation of swallowing (Fees®) using the Penetration–Aspiration Scale: a replication study. Dysphagia. 2002;17(4):308-15.

84. Kelly AM, Drinnan MJ, Leslie P. Assessing penetration and aspiration: how do videofluoroscopy and fiberoptic endoscopic evaluation of swallowing compare? The Laryngoscope. 2007;117(10):1723-7.

85. Gullung JL, Hill EG, Castell DO, Martin-Harris B. Oropharyngeal and Esophageal Swallowing Impairments: Their Association and the Predictive Value of the Modified Barium Swallow Impairment Profile and Combined Multichannel Intraluminal Impedance—Esophageal Manometry. Annals of Otology, Rhinology & Laryngology. 2012;121(11):738-45.

86. Kim DH, Choi KH, Kim HM, Koo JH, Kim BR, Kim TW, et al. Inter-rater reliability of videofluoroscopic dysphagia scale. Annals of Rehabilitation Medicine. 2012;36(6):791-6.

87. Kim J, Oh B-M, Kim JY, Lee GJ, Lee SA, Han TR. Validation of the videofluoroscopic dysphagia scale in various etiologies. Dysphagia. 2014;29(4):438-43.

88. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Medical Research Methodology. 2010;10(1):22.

89. Logemann JA. Manual for the videofluorographic study of swallowing: Pro ed; 1993.

90. Ickenstein GW, Höhlig C, Prosiegel M, Koch H, Dziewas R, Bodechtel U, et al. Prediction of outcome in neurogenic oropharyngeal dysphagia within 72 hours of acute stroke. Journal of Stroke and Cerebrovascular Diseases. 2012;21(7):569-76.

91. van der Maarel-Wierink CD, Vanobbergen JN, Bronkhorst EM, Schols JM, de Baat C. Meta-analysis of dysphagia and aspiration pneumonia in frail elders. Journal of Dental Research. 2011;90(12):1398-404.

92. Streiner DL, Norman GR, Cairney J. Item Response Theory. Health measurement scales: a practical guide to their development and use: Oxford University Press, USA; 2014.

93. Fan X. Item response theory and classical test theory: an empirical compraison of their item/person statistics. Educational and Psychological Measurement. 1998;58(3):357.

94. Duong M. Introduction to Item Response Theory and Its Applications. Michigan State University, 2004.

95. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research. 2007;16(5).

96. Reise SP, Ainsworth AT, Haviland MG. Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research. Current Directions in Psychological Science. 2005;14(2):95-101.

[Supplementary table 1 here]

Table 1: Search Terms

	Database and Search Terms	Limits	Number of records
Subject Heading	**CINAHL:** ((MH "Outcome Assessment") OR (MH "Patient Assessment") OR (MH "Clinical Assessment Tools") OR (MH "Speech and Language Assessment") OR (MH "Health Impact Assessment") OR (MH "Needs Assessment") OR (MH "Functional Assessment") OR (MH "Self Assessment") OR (MH "Physical Examination") OR (MH "Functional Assessment Inventory") OR (MH "Measurement Issues and Assessments") OR (MH "Neurologic Examination") OR (MH "Weights and Measures") OR (MH "Behavior Rating Scales") OR (MH "Questionnaires") OR (MH "Scales") OR (MH "Clinical Assessment Tools") OR (MH "Health Screening") OR (MH "Outcome Assessment") OR (MH "Evaluation") OR (MH "Disability Evaluation") OR (MH "Health Status") OR (MH "Health Status Indicators") OR (MH "Neurologic Examination") OR (MH "Physical Examination") OR (MH "Research Instruments") OR (MH "Research Measurement")) AND ((MH "Deglutition Disorders") OR (MH "Deglutition") OR (MH "Infant Feeding") OR (MH "Feeding and Eating Disorders of Childhood") OR (MH "Feeding of Disabled") OR (MH "Infant Feeding, Supplemental") OR (MH "Eating Behavior") OR (MH "Eating") OR (MH "Eating Behavior") OR (MH "Eating Disorders")) AND ((MH "Psychometrics") OR (MH "Measurement Issues and Assessments") OR (MH "Validity") OR (MH "Predictive Validity") OR (MH "Reliability and Validity") OR (MH "Internal Validity") OR (MH "Face Validity") OR (MH "External Validity") OR (MH "Discriminant Validity") OR (MH "Criterion-Related Validity") OR (MH "Consensual Validity") OR (MH "Concurrent Validity") OR (MH "Qualitative Validity") OR (MH "Construct Validity") OR (MH "Content Validity") OR (MH "Instrument Validation") OR (MH "Validation Studies") OR (MH "Test-Retest Reliability") OR (MH "Sensitivity and Specificity") OR (MH "Reproducibility of Results") OR (MH "Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR (MH "Measurement Error") OR (MH "Bias (Research)") OR (MH "Selection Bias") OR (MH "Sampling Bias") OR (MH "Precision") OR (MH "Sample Size Determination") OR (MH "Repeated Measures")) AND ((MH "Endoscopy") OR (MH "Endoscopy, Gastrointestinal") OR (MH "Endosonography") OR (MH "Endoscopy, Digestive System") OR (MH "Endosonography") OR (MH "Fluoroscopy") OR (MH "Radiography") OR (MH "Radiography, Dental") OR (MH "Radiography, Bitewing") OR (MH "Tomography, X-Ray") OR (MH "Radiography, Computed") OR (MH "Radiography, Interventional") OR (MH "Radiography, Dental, Digital") OR (MH "Radiography, Thoracic") OR (MH "Radiography, Panoramic") OR (MH "Neuroradiography") OR (MH "Esophagoscopy") OR (MH "Manometry") OR (MH "Electric Impedance") OR (MH "Electrical Stimulation, Functional") OR (MH "Electromyography") OR (MH "Neural Conduction") OR (MH "Radionuclide Imaging") OR (MH "Tomography, Spiral Computed") OR (MH "Diagnostic Imaging") OR (MH "Tomography") OR (MH "Multidetector Computed Tomography") OR (MH "Tomography, X-Ray Computed") OR (MH "Tomography, Emission-Computed") OR (MH "Tomography, Emission-Computed, Single-Photon") OR (MH "Tomography, X-Ray") OR (MH "Tomography, Optical Coherence") OR (MH "Tomography, Optical") OR (MH "Computed Tomography Angiography") OR (MH "Ultrasonography") OR (MH "Ultrasonics") OR (MH "Kinesiology") OR (MH "Kymography") OR (MH "Electrokymography"))	NA	94

	Embase: (measurement/ OR diagnostic procedure/ OR rating scale/ OR screening/ OR screening test/ OR questionnaire/ OR outcome assessment/ OR evaluation study/ OR medical informatics/ OR health status/ OR examination/ OR diagnostic procedure/ OR diagnostic test/ OR diagnostic approach route/) AND (dysphagia/ OR swallowing/ OR feeding/ OR feeding behavior/ OR feeding disorder/ OR feeding difficulty/ OR eating/ OR eating disorder/) AND (psychometry/ OR validity/ OR reliability/ OR measurement error/ OR measurement precision/ OR measurement repeatability/ OR error/ OR statistical bias/ OR test retest reliability/ OR intrarater reliability/ OR interrater reliability/ OR accuracy/ OR criterion validity/ OR internal validity/ OR face validity/ OR external validity/ OR discriminant validity/ OR concurrent validity/ OR qualitative validity/ OR construct validity/ OR content validity/) AND (endoscopy/ OR fiberscope endoscopy/ OR videoendoscopy/ OR high resolution endoscopy/ OR endoscopic ultrasonography/ OR radiography/ OR fluoroscopy/ OR esophagography/ OR esophagoscopy/ OR manometry/ OR impedance/ OR esophageal manometer/ OR electromyogram/ OR electromyography/ OR nerve conduction/ OR scintigraphy/ OR bone scintiscanning/ OR scintiscanning/ OR tomography/ OR diagnostic imaging/ OR ultrasound/ OR echography/ OR kinematics/ or kinesiology/ OR kymography/ OR electrokymography/ OR laryngography/)	NA	119
	Medline: ("Weights and Measures"/ OR Mass Screening/ OR "Surveys and Questionnaires"/OR "Outcome Assessment (Health Care)"/ OR Evaluation Studies as Topic/ OR health status/ OR Health Status Indicators/) AND (Deglutition Disorders/ OR Deglutition/ OR Feeding and Eating Disorders/ OR feeding behavior/ OR Eating/) AND (psychometrics/ OR "Bias (Epidemiology)"/) AND ((endoscopy/ OR Endosonography/ OR Radiography/ OR Fluoroscopy/ OR esophagoscopy/ OR Manometry/ OR Electric Impedance/ OR electromyography/ OR Neural conduction/ OR Radionuclide Imaging/ OR tomography/ OR Diagnostic imaging/ OR ultrasonography/ OR ultrasonics/ OR Biomechanical Phenomena/ OR kymography/ OR electrokymography/) OR (FEES OR FEEST OR VFS OR VFSS OR MBS OR (barium AND swallow*) OR endoscop* OR videoendoscop* OR video-endoscop* OR naso-endoscop* OR nasoendoscop* OR videofluoroscop* OR fluoroscop* OR radiogra* OR imag* OR neuroradiogr* OR pneumoradiogra* OR endosonogra* OR esophagoscop* OR esophagogra* OR HRM OR manomet* OR videomanomet* OR impedanc* OR bioimpedanc* OR plethysmogra* OR electromyogra* OR EMG OR sEMG OR electric* OR (neural AND conduction) OR (nerve AND conduction) OR scintigra* OR scintiscan* OR (bone AND scan*) OR tomogra* OR X-ray* OR ultraso* OR sonogr* OR kinesiolog* OR biomechanic* OR kinematic* OR EGG OR electroglottogra* OR kymogram* OR videokymogra* OR electrokymogra* OR (high AND speed AND recording) OR (high-speed AND recording)))		130

	PubMed: ("Outcome and Process Assessment (Health Care)"[Mesh] OR "Needs Assessment"[Mesh] OR "Patient Outcome Assessment"[Mesh] OR "Symptom Assessment"[Mesh] OR "Outcome Assessment (Health Care)"[Mesh] OR "Self-Assessment"[Mesh] OR "Patient Acuity"[Mesh] OR "Neurologic Examination"[Mesh] OR "Diagnosis"[Mesh] OR "Weights and Measures"[Mesh] OR "Severity of Illness Index"[Mesh] OR "Neuropsychological Tests"[Mesh] OR "Behavior Rating Scale"[Mesh] OR "Visual Analog Scale"[Mesh] OR "diagnosis" [Subheading] OR "Surveys and Questionnaires"[Mesh] OR "Treatment Outcome"[Mesh] OR "Patient Reported Outcome Measures"[Mesh] OR "Diagnostic Self Evaluation"[Mesh] OR "Disability Evaluation"[Mesh] OR "Health Status"[Mesh] OR "Health Status Indicators"[Mesh] OR "Physical Examination"[Mesh]) AND ("Deglutition Disorders"[Mesh] OR "Deglutition"[Mesh] OR "Feeding and Eating Disorders of Childhood"[Mesh] OR "Feeding Behavior"[Mesh] OR "Feeding and Eating Disorders"[Mesh] OR "Eating"[Mesh]) AND ("Psychometrics"[Mesh] OR "Reproducibility of Results"[Mesh] OR "Validation Studies as Topic"[Mesh] OR "Validation Studies" [Publication Type] OR "Bias (Epidemiology)"[Mesh] OR "Observer Variation"[Mesh] OR "Selection Bias"[Mesh] OR "Diagnostic Errors"[Mesh] OR "Dimensional Measurement Accuracy"[Mesh] OR “Predictive Value of Tests"[Mesh] OR "Discriminant Analysis"[Mesh]) AND ("Endoscopy"[Mesh] OR "Endoscopy, Digestive System"[Mesh] OR "Endosonography"[Mesh] OR "Radiography"[Mesh] OR "Diagnostic Imaging"[Mesh] OR "Radiography, Dental, Digital"[Mesh] OR "Radiography, Bitewing"[Mesh] OR "Radiography, Dual-Energy Scanned Projection"[Mesh] OR "Radiography, Thoracic"[Mesh] OR "Radiography, Interventional"[Mesh] OR "Radiography, Panoramic"[Mesh] OR "Radiography, Dental"[Mesh] OR "Pneumoradiography"[Mesh] OR "Fluoroscopy"[Mesh] OR "Esophagoscopy"[Mesh] OR "Manometry"[Mesh] OR "Electric Impedance"[Mesh] OR "Plethysmography, Impedance"[Mesh] OR "Electromyography"[Mesh] OR "Neural Conduction"[Mesh] OR "Radionuclide Imaging"[Mesh] OR "Tomography, Emission-Computed"[Mesh] OR "Tomography, X-Ray Computed"[Mesh] OR "Tomography"[Mesh] OR "Single Photon Emission Computed Tomography Computed Tomography"[Mesh] OR "Positron Emission Tomography Computed Tomography"[Mesh] OR "Multidetector Computed Tomography"[Mesh] OR "Four-Dimensional Computed Tomography"[Mesh] OR "Electron Microscope Tomography"[Mesh] OR "Spiral Cone-Beam Computed Tomography"[Mesh] OR "Cone-Beam Computed Tomography"[Mesh] OR "Tomography, Optical Coherence"[Mesh] OR "Positron-Emission Tomography"[Mesh] OR "Tomography, Spiral Computed"[Mesh] OR "Tomography Scanners, X-Ray Computed"[Mesh] OR "Tomography, X-Ray"[Mesh] OR "Tomography, Emission-Computed"[Mesh] OR "Computed Tomography Angiography"[Mesh] OR "X-Ray Microtomography"[Mesh] OR "Echo-Planar Imaging"[Mesh] OR "Magnetic Resonance Imaging"[Mesh] OR "Ultrasonography"[Mesh] OR "Diagnostic Imaging"[Mesh] OR "Ultrasonics"[Mesh] OR "Biomechanical Phenomena"[Mesh] OR "Kinesiology, Applied"[Mesh] OR ”Kymography"[Mesh] OR "Electrokymography"[Mesh])	NA	1,287

Free Text	**CINAHL:** (assessment* OR measure* OR questionnaire* OR test OR tests OR scale* OR screening* OR evaluation) AND (dysphag* OR swallowing* OR deglut* OR feed* OR eat*) AND (psychometric* OR reliability* OR validit* OR reproducibility* OR bias OR responsiveness) AND (FEES OR FEEST OR VFS OR VFSS OR MBS OR (barium AND swallow*) OR endoscop* OR videoendoscop* OR video-endoscop* OR naso-endoscop* OR nasoendoscop* OR videofluoroscop* OR fluoroscop* OR radiogra* OR imag* OR neuroradiogr* OR pneumoradiogra* OR endosonogra* OR esophagoscop* OR esophagogra* OR HRM OR manomet* OR videomanomet* OR impedanc* OR bioimpedanc* OR plethysmogra* OR electromyogra* OR EMG OR sEMG OR electric* OR (neural AND conduction) OR (nerve AND conduction) OR scintigra* OR scintiscan* OR (bone AND scan*) OR tomogra* OR X-ray* OR ultraso* OR sonogr* OR kinesiolog* OR biomechanic* OR kinematic* OR EGG OR electroglottogra* OR kymogram* OR videokymogra* OR electrokymogra* OR (high AND speed AND recording) OR (high-speed AND recording))	Search fields: Title and/or Abstract Publication date: 01/02/2016 to 08/02/2017	14
	Medline: As per CINAHL Free Text	Search fields: Title and/or Abstract Publication date: “2016 – Current”	125
	Embase: As per CINAHL Free Text	Search fields: Title and/or Abstract Publication date: “2016 – Current”	179
	PubMed: As per CINAHL Free Text	Search fields: Title and/or Abstract Publication date: from 08/02/2016 to 08/02/2017	142

Table 2: COSMIN definitions of domains, psychometric properties and aspects of psychometric properties for Health-Related Patient-Reported Outcomes adapted from Mokkink et al. (35)

	Psychometric Property	Domain
Definition[a]		**Validity** **The extent to which an instrument measures the construct/s that it claims to measure**
	Content Validity	The degree that the content of an instrument adequately reflects the construct to be measured (includes face validity)
	Face validity[b]	The degree to which instrument (items) appear to be an adequate reflection of the construct to be measured
	Construct Validity	The extent to which the scores of an instrument are consistent with hypotheses, based on the assumption that the instrument is a valid measure of the construct being measured
	Structural validity[c]	The extent to which instrument scores adequately reflect the dimensionality of the construct to be measured
	Hypothesis testing[c]	Item construct validity
	Cross cultural validity[c]	The degree to which the performance of items on a translated or culturally adapted measure are an adequate reflection of the performance of the items in the original version
	Criterion Validity	The degree to which the scores of an instrument satisfactorily reflect a 'gold standard'
		Reliability **The degree to which the measurement is free from measurement error**
	Internal Consistency	The level of correlation amongst items
	Reliability	The proportion of total variance in the measurements due to "true" differences amongst patients
	Measurement Error	The error of a patient's score, systematic and random, not attributed to true changes in the construct measured
		Responsiveness **The capability of an HR-PRO instrument to detect change in the construct to be measured over time**
	Responsiveness	Item responsiveness
		Interpretability[d] **The extent to which qualitative meaning can be given to an instrument's quantitative scores or score change**

Notes

[a] Applies to Health-Related Patient-Reported Outcomes (HR-PRO) instruments.

[b] Aspect of content validity under the domain of validity.

[c] Aspects of construct validity under the domain of validity.

[d] Interpretability is not considered a psychometric property

Table 3: Criteria of psychometric quality rating based on Terwee et al.(36) and Schellingerhout et al. (38)

Psychometric Property	Score[a]	Quality Criteria[b]
Content Validity	+	A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection and target population and (investigators or experts) were involved in item selection
	?	A clear description of above-mentioned aspects is lacking or only target population involved or doubtful design or method
	-	No target population involvement
	±	Conflicting results
	NR	No information found on target population involvement
	NE	Not evaluated
Structural validity[c]	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	-	Factors explain <50% of the variance
	±	Conflicting results
	NR	No information found on structural validity

Psychometric Property	Score[a]	Quality Criteria[b]
	NE	Not evaluated
Hypothesis testing[c]	+	Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses
	?	Doubtful design or method (e.g., no hypotheses)
	-	Less than 75% of hypotheses were confirmed, despite adequate design and methods
	±	Conflicting results between studies within the same manual
	NR	No information found on hypotheses testing
	NE	Not evaluated
Internal consistency	+	Factor analyses performed on adequate sample size (7 * # items consistency and ≥100) AND Cronbach's alpha(s) calculated per dimension and Cronbach's alpha(s) between 0.70 and 0.95
	?	No factor analysis OR doubtful design or method
	-	Cronbach's alpha(s) <0.70 or >0.95, despite adequate design and method
	±	Conflicting results
	NR	No information found on internal consistency
	NE	Not evaluated
Reliability	+	ICC or weighted Kappa ≥0.70
	?	Doubtful design or method (e.g., time interval not mentioned)
	-	ICC or weighted Kappa < 0.70, despite adequate design and method
	±	Conflicting results
	NR	No information found on reliability
	NE	Not evaluated
Measurement error[d]	+	MIC < SDC OR MIC outside the LOA OR convincing arguments that agreement is acceptable
	?	Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement is acceptable)
	-	MIC ≥ SDC OR MIC equals or inside LOA, despite adequate design and method;
	±	Conflicting results
	NR	No information found on measurement error
	NE	Not evaluated

[a]Scores: + = positive rating, ? = indeterminate rating, — = negative rating, ± = conflicting data, NR = not reported, NE = not evaluated (for study of poor methodological quality according to COSMIN rating, data are excluded from further evaluation).

[b]Doubtful design or method is assigned when a clear description of the design or methods of the study is lacking, sample size smaller than 50 subjects (should be at least 50 in every subgroup analysis), or any important methodological weakness in the design or execution of the study

[c]Hypothesis testing: all correlations should be statistically significant (if not, these hypotheses are not confirmed) AND these correlations should be at least moderate (r > 0.5)

[d]Measurement error: MIC = minimal important change, SDC = smallest detectable change, LOA = limits of agreement.

Table 4:** **Revised criteria for levels of evidence for the overall quality of the measurement properties based on Schellingerhout et al. (38)

Level	Criteria
Strong	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	Consistent findings in multiples studies of fair methodological quality OR in one study of good methodological quality
Limited	One study of fair methodological quality
Conflicting	Conflicting findings
Not Evaluated[a]	Only studies of poor methodological rating (COSMIN)
Indeterminate[b]	Only indeterminate data on measurement properties

[a] Not evaluated = only studies of poor methodological quality according to COSMIN; data from these studies are excluded from further analyses.

[b] Indeterminate = only indeterminate outcome data on the assessment measurement property (score: '?'), therefore, also indeterminate level of evidence for the overall quality of that measurement property

Table 5: Methodological quality assessment of studies reporting on reliability only (COSMIN (27), quality of reliability per study (criteria by Terwee et al. (36) and Schellingerhout et al. (38) and overall quality score for reliability per measure (Schellingerhout et al. (38)

					Reliability[b]		
Measure; Reference, Year published	**Study on psychometrics**	**Aspects evaluated by measure**	**Total number of items[a]; Domain of variables**	**Response options**	**COSMIN quality score**	**Quality of psychometric properties and - rater reliability**	**Overall quality score**
			FEES				
Unnamed Marvin et al. (44), 2016	Marvin et al. (44)	Presence of secretions, location of sections, colour of secretions and airway invasion (penetration / aspiration) differentiated by bolus dye colours (green or white)	4 Volume and spatial	Nominal scales describing impairment; e.g. 'colour: clear, white, brown, yellow or bloody'	Fair (42.42%)	**Inter: NR** **Intra:** **Using green bolus: +** **Using white bolus: -**	Limited (positive) Limited (negative)
Unnamed Pilz et al. (50), 2016[c]	Pilz et al. (50)	Piecemeal deglutition (number of swallows on same bolus), residue in pyriform and valleculae and laryngeal penetration / aspiration	4 Volume, spatial and patient response	Ordinal rating scales ranging from 3 to 5-points; e.g. 'bolus retention in the valleculae after swallowing: 0 = no pooling, 1 = filling of <50% of the vallculae, 2 = filling of >50% of valleculae'	Excellent (78.79%)	**Inter**: ± **Intra**: **+**	Conflicting
Unnamed Rodiriguez et al. (55), 2007	Rodiriguez et al. (55)	Adequacy of pharyngeal wall movement and ability to complete a swallow maneuverer (pharyngeal squeeze)	2 Spatial	Pharyngeal wall movement: 3 option nominal scale ('normal', 'diminished' or 'absent') Pharyngeal squeeze maneuverer: dichotomous scale ('normal' or 'abnormal')	Fair (48.48%)	**Inter**:**?** **Intra**: **+**	Indeterminate
Unnamed[c] Susa et al. (48), 2015	Susa et al. (48)	Pattern of soft palate movement during continuous drinking via a straw	1 Temporal	Nominal. Raters selected one descriptor which best described swallow: V- segmental (velopharynx opens post swallow); V- continuous (velopharynx closure continues after swallow); Or V-mixed (both V-segmental and V mixed swallows present).	Fair (42.42%)	**Inter**: **+** **Intra**: **+**	Limited (positive)

Physiological and swallowing evaluation form[c] Tohara et al. (56), 2010	Tohara et al. (56)	Physiological evaluation: describes anatomical location of secretions, contraction of pharyngeal wall, glottal closure. Swallow evaluation: notes premature spillage, swallow reflex onset, condition of masticated food, bolus formation, whiteout, aspiration (including type, amount and depth), cough after aspiration, valleculae residue, pyriform sinus residue, pharyngeal wall residue	16 Volume, temporal, spatial and patient response	Nominal and ordinal scales with between three and eight descriptors; e.g. 'aspiration type: prior, during, after'	Good (63.63%)	**Inter: -** **Intra: ±**	Conflicting
Unnamed Warnecke et al. (49), 2016	Warnecke et al. (49)	Premature spillage, penetration / aspiration and residue	3 Volume, spatial and patient response	Ordinal scales with five levels; e.g. 'premature spillage: 0 – the bolus is behind the tongue … 4 – the bolus falls into the laryngeal vestibule'	Good (57.58%)	**Inter**: **+** **Intra**: **+**	Moderate (positive)
		VFSS					
Unnamed Bryant et al. (57), 2012	Bryant et al. (57)	Bolus holding, bolus formation, lip closure, poor bolus control, piecemeal deglutition, prolonged oral transit time, oral stasis, poor tongue coordination, pharyngeal delay, prolonged transit time, laryngeal elevation, velar elevation, vallecular stasis, pyriform sinus retention, reduced pharyngeal wall contraction, reduced epiglottic movement, reduced swallow respiratory coordination, dilation, reflux, Zenker's diverticulum, degree of aspiration, degree of penetration	23 Volume, temporal, spatial and patient response	5-point ordinal scale for all items, ranging from 0 (not observed) to 4 (severe impairment), with the exception of Zenker's diverticulum and aspiration/penetration. Nominal scale for Zenker's ('not observed', 'yes', 'no'), and aspiration/penetration ('not observed', 'mild', 'moderate' 'severe') <u>Note</u>: Reliability analysed for the following aspects only: Impaired base of tongue function, pharyngeal delay, impaired pharyngeal wall contraction, impaired laryngeal function, impaired epiglottic function, impaired UES function	Good (52.63%)	**Inter**: **-** **Intra: NR**	Moderate (negative)
Bethlehem Assessment Scale (BAS) Scott (70), 1999	Frowen et al. (23)	Describes severity of impairment or identifies normal function of eleven features of the swallow act (lip, tongue and function, velum elevation, swallow reflex, hyoid elevation, valleculae and pyriform residue, aspiration and pharyngeal wall and cricopharyngeal function)	11 Volume, temporal and spatial	4-point ordinal scale (1 – 4) with corresponding descriptors from 'normal' to 'severe dysfunction'	Good (57.6%)	**Inter**: **+** **Intra**: **+**	Moderate (positive)

Unnamed Gibson et al. (52) 1995	Gibson et al. (52)	Aspiration, oral and pharyngeal duration time, number of swallows required to clear pharynx of the bolus, number of posterior tongue elevations per bolus, place of bolus initiation of the swallow and valleculae pooling post-swallow	6 Volume, temporal and spatial	Open-ended response options for continuous variables (e.g. time in seconds of pharyngeal phase) and nominal scales with 3 descriptors (e.g. 'amount of residue: whole part or none') for other variables	Fair (50.00%)	**Inter**: ± **Intra**: ±	Conflicting
Temporal and Physiologic Features of Infant Swallows Gosa et al. (41), 2015	Gosa et al. (41)	Describes number of sucks per swallow, suck and oral transit time, velar movement, collection of bolus pre-swallow, pharyngeal transit time, duration cricopharyngeal opening / pharyngeal constriction and laryngeal closure, time to complete laryngeal closure, epiglottic tilting, nasopharyngeal backflow, penetration / aspiration, residue and jaw position	16 Volume, temporal, spatial and patient response	Nine continuous variables (time measured in seconds and number of downward motions of mandible). Three nominal scales; e.g. 'jaw position – open, closing, neutral' Four ordinal scales; e.g. 'epiglottic tilting: yes / no'	Fair (41.67%)	**Inter**: **?** **Intra**: **?**	Indeterminate
Unnamed Hind et al. (42), 2009	Hind et al. (42)	Presence or absence of aspiration	1 Spatial	Dichotomous options of presence / absence of aspiration	Good (52.63%)	**Inter: +** **Intra: NR**	Moderate (positive)
'Objective measures' based on norms from Leonard et al. (71) Lee et al. (53)	Lee et al. (53)	Hyoid elevation, pharyngeal area, pharyngeal constriction ratio and pharyngo-oesophageal segment opening	4 Spatial	Dichotomous options of normal / abnormal	Good (54.55%)	**Inter: -** **Intra**: ±	Conflicting
Unnamed Mann et al. (43), 2000	Mann et al. (43)	Oral preparation (forming and holding bolus), oral transit time, pharyngeal phase (triggering of swallow, motion of pharyngeal anatomy, movement and management of bolus through pharynx) and aspiration	7 variables describing swallow 2 variables indicating overall diagnosis. Volume, temporal, spatial and patient response	Continuous measures of duration (e.g. time from arrival of bolus head at mandible ramus until tail passes oesophageal sphincter), estimates of volume and frequency (e.g. amount and frequency of aspiration) and range of motion (e.g. hyoid movement). Overall impression: Two 5-point nominal scales of dysphagia and aspiration (e.g. normal, mild, moderate, severe, complete) with criterion at each point	Good (68.42%)	**Inter: Diagnosis of dysphagia: +** **Diagnosis of aspiration: -** **Intra: NR**	Moderate (positive) Moderate (negative)

Unnamed Miles (54), 2016	Miles (54)	Oesophageal features: bolus transit, stasis, level of stasis, redirection, and if onwards referral to a specialist is required	5 Temporal and spatial	Dichotomous options; e.g. stasis: present / absent. Referral required: Yes / No	Fair (36.84%)	**Inter: ±** **Intra: NR**	Conflicting
Unnamed McCullough et al. (72),1999	McCullough et al. (26)	Lingual control, oral, vallecular, pyriform and hypopharyngeal residue, epiglottic function, hyolaryngeal excursion, cricopharyngeal prominence, oral and pharyngeal transit duration, total swallow duration, pharyngeal delay time and duration upper oesophageal sphincter opening	13 Volume, temporal and spatial	Oropharyngeal function: Dichotomous options; e.g. lingual control: considered present if evidence of reduced lingual propulsion of the bolus Open-ended questions on duration measures: time of events in relation to bolus movements and anatomical movements	Fair (48.48%)	**Inter: -** **Intra:** ±	Conflicting
'VFSS objective measures' adapted from Leonard and Kendall (73),1997	Nordin et al. (45)	Total pharyngeal transit time, airway closure duration, pharyngeal - oesophageal opening duration, maximum pharyngeal constriction, pharyngeal constriction ratio, pharyngeal - oesophageal maximum opening width	5 Volume and spatial	Open-ended options, with instructions on how to calculate duration / space utilised; e.g. pharyngeal - oesophageal opening duration- rater subtracts time when upper oesophageal sphincter opens from time when it closes to calculate total duration	Good (60.0%)	**Inter: +** (note: '+' score achieved only following 8 weeks of training. Initially all '-') **Intra: NR**	Moderate (positive)
Unnamed Power et al. (46), 2009	Power et al. (46)	Oral transit time, pharyngeal transit time, swallow response time, laryngeal closure duration, cricopharyngeal opening duration	5 Temporal	Open-ended options, with instructions on how to calculate duration. Raters reported in continuous measure (seconds)	Good (60.0%)	**Intra: +**	Moderate (positive)
Bolus residue scale[c] Rommel et al. (47), 2015	Rommel et al. (47)	Spread of pharyngeal residue with reference to anatomical structures affected	1 Spatial	6-point ordinal scale with descriptors at each level; e.g. '1 – no residue … 6 – residue in vallecullae and posterior pharyngeal wall and pyriform sinus'	Fair (33.33%)	**Inter: +** **Intra: +**	Limited (positive)
Modified Charing Cross Hospital Dysphagia Profile Unknown, 1998	Scott et al. (59)	Lip, tongue and jaw function, velar, hyoid, pharyngeal wall and cricopharyngeal movement, valleculae and pyriform residue and presence of aspiration	11 Volume, temporal, spatial and patient response	5-point ordinal scale with descriptors at each level; e.g. 'tongue function: 1 – bolus is propelled completely into pharynx in a smooth, uninterrupted wave-like motion'	Poor (15.79%)	**Inter: NE** **Intra: NR**	NE
Unnamed[c] Stoeckli et al. (58), 2003	Stoeckli et al. (58)	Lip closure, soft palate / tongue back seal, bolus transport / lingual motion, delayed initiation, soft palate elevation, tongue base retraction, laryngeal elevation, laryngeal closure,	16 Volume, temporal, spatial and patient response	8 – point ordinal scale to describe depth of penetration / aspiration and patient response. Variety of nominal scales with two to six descriptors for remaining variables; e.g. 'Lip closure: insufficient / sufficient	Fair (47.37%)	**Inter: -** **Intra: NR**	Limited (negative)

		anterior hyolaryngeal excursion, pharyngeal contraction, upper oesophageal sphincter opening / closure, penetration / aspiration, residue		Residue location: floor of mouth, base of tongue, valleculae, pharyngeal wall, aryepiglottic folds, pyriform sinuses'			
FEES and VFSS							
Pharyngeal Residue Severity Scale Kelly et al. (51), 2006	Kelly et al. (51)	Volume of pharyngeal residue	1 Volume	Nominal scale reporting volume of residue: 'none', 'coating', 'mild', 'moderate' or 'severe'	Good (63.64%)	**Inter**: - **Intra**: +	Conflicting

Notes:

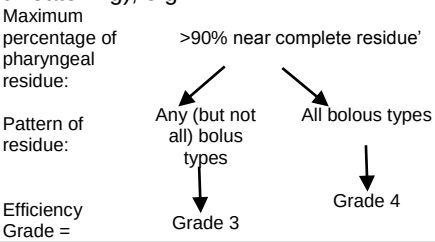
[a] **Items:** the list of variables the measure seeks to assess, such as oral transit time or pyriform residue. A single item may attempt to assess multiple features of the variable (e.g. the item 'severity of aspiration' may assess volume of aspirate, spatial distance of aspirate, time when aspiration occurred and patient's response to aspiration event).

[b] **COSMIN quality score:** The quality of the studies that evaluated the psychometric properties of each instrument was evaluated according to the COSMIN rating per item: four-point scale was used (1 = Poor, 2 = Fair, 3 = Good, 4 = Excellent). The overall methodological quality per study was presented as percentage of rating (Poor = 0–25.0%, Fair = 25.1%-50.0%, Good = 50.1%-75.0%, Excellent = 75.1%-100.0%) NR: not reported
Quality of psychometric properties: based on the criteria by Terwee et al. (36) and Schellingerhout (38) (see Table 3)
Overall quality score: combined COSMIN methodological quality and Terwee et al. (36) and Schellingerhout (38) (see Table 4)

[c] **Measure likely created in language other than English**. Attempt to contact all authors; no information available on translation process, with the exception of Pilz et al. (50). Pilz et al. (50) reported the measure was originally created in Dutch, then subsequently translated to English using a professional translator. Translation process score according to COSMIN: 33.33% (Fair)

Table 6: Descript ion of measures with multiple known psychometric properties

Measure; Reference, Year Published	Aspects evaluated by measure	Summed score / number of subscales[a]	Total number of items; domain of variables	Response Options
		FEES		
Marionjoy 3-Point secretion severity scale (69), 2003	Volume of secretions present	Nil summed score; nil subscales	1 Volume	3-point ordinal scales with descriptors corresponding to each score; 'functional', 'severe' or 'profound' Definitions provided for each descriptor: e.g. 3 = 'profound – secretions present on vocal cords and / or tracheal aspiration of secretions'
Marionjoy 5 – Point Secretion Severity Scale (69), 2003	Volume of secretions present	Nil summed score; nil subscales	1 Volume	5-point ordinal scales with descriptors at each score; 'normal', 'mild', 'moderate', 'severe', or 'profound' Definitions provided for each descriptor: e.g. '2= mild – pooling of pharyngeal secretions from 10% - 25% in pyriform sinuses and / or vallecular space'
Dysphagia Score **(74)**, 2008	Presence or absence of secretions, residue and protective airway reflexes	Nil summed score; nil subscales	1 – 4 (increasingly challenging bolus textures) Volume, spatial and patient response	Ordinal 6-point scale with descriptors at each score describing symptoms; e.g. 'Liquids – penetration without or insufficient protective reflex' Scores dependent on patient performance at level of bolus challenge (e.g. puree up to soft solid food)
Pooling-Score (P-Score) (60), 2008	Anatomical site of residue, volume of residue and number of swallows required to clear residue	Summed score, three subscales (site, amount, management)	3 Volume, spatial and patient response	Nominal scale, with a score assigned to each descriptor (endoscopic landmark) within each subscale. Raters choose one descriptor only per subscale. Subscales then summed
Boston Residue and Clearance Scale (BRACS) (61), 2013	Amount and location of pharyngeal residue and patient's ability to clear residue	Single overall summed score, nil subscales	16 Volume, spatial and patient response	Ordinal 4-point scales (0 – 3) with severity descriptors (none – severe). Scoring completed across four anatomical 'zones', comprised of 12 sites in the laryngopharynx. Four additional options for if residue in four or more regions - residue presence / absence in vestibule and presence / absence / effectiveness of clearing swallows
Yale Pharyngeal Residue Severity Rating Scale **(75)**, 2015	Residue in pharynx	Nil summed score; nil subscales	2 Volume and spatial	5-point ordinal scale with descriptors corresponding to each score; e.g. 'Trace: 1 – 5%, trace coating of the mucosa'
Murray Secretion Severity Rating Scale (Secretion Scale) **(76)**, 1996	Secretions in hypo-pharynx in terms of location, volume and patient response	Nil summed score; nil subscales	1 Volume, spatial and patient response	Ordinal 4-point scales (0 – 3) with verbal descriptors; e.g. '0 – most normal rating. No visible secretions anywhere in hypopharynx or some transient bubbles visible in the valleculae and pyriform sinuses. Those secretions were not bilateral or deeply pooled'
		VFSS		
Modified Barium Swallowing Study swallowing evaluation tool (MBSImp) (62), 2008	Lip closure, bolus hold position / tongue control, bolus preparation / mastication, bolus transport / lingual motion, oral residue, initiation of the pharyngeal swallow, soft palate elevation, laryngeal elevation, anterior hyoid motion, epiglottic movement, laryngeal closure, pharyngeal	Nil summed score; seventeen 'components' which are individually rated for each bolus texture	17 Volume, temporal, spatial and patient response	3 to 5-point ordinal scales, with verbal descriptors at each score; e.g. component 6, initiation of pharyngeal swallow: '0= bolus head at posterior angle of ramus 1= Bolus head at vallecular pit 2= bolus head at posterior laryngeal surface of epiglottis'

Measure; Reference, Year Published	Aspects evaluated by measure	Summed score / number of subscales[a]	Total number of items; domain of variables	Response Options
	stripping wave, pharyngeal contraction, cricopharyngeal opening, tongue base retraction, pharyngeal residue and oesophageal clearance			'Overall impression' score per swallow component also applied, which derives from scores across multiple bolus presentations
Functional Dysphagia Scale (FDS) (63), 2001	Lip closure, bolus formation, residue in oral cavity, oral transit time, triggering pharyngeal swallow, laryngeal elevation and epiglottis closure, nasal penetration, residue in valleculae, residue in pyriform sinus, coating of pharyngeal wall after swallow, pharyngeal transit time	Variables have associated numerical scores which are summed to create 'total score'; nil subscales	11 Volume, temporal, and spatial	Nominal scales, with values which vary between variables; e.g. 'lip closure: intact, inadequate, none. Residue in oral cavity: none, <10%, 10-50%, >50%' Each value has an associated numerical score, ranging from 0 to 12
Video-fluroscopic Dysphagia Scale (VDS) (64), 2008	Lip closure, bolus formation, mastication, apraxia, tongue to palate contact, premature bolus loss, oral transit time, triggering pharyngeal swallow, vallecular residue, laryngeal elevation, pyriform sinus residue, coating of pharyngeal wall, pharyngeal transit time, aspiration	Variables have associated numerical scores which are summed to create a 'total score'; nil subscales	14 Volume, temporal, and spatial	Nominal scales, with values which vary between variables; e.g. 'lip closure: intact, inadequate, none. Premature bolus loss: none, <10%, 10-50%, >50%' Each value has an associated numerical score ranging from 0 to 13.5
Dynamic Imaging Grade of Swallowing Toxicity Scale (DIGEST) (66), 2017	Penetration, aspiration and pharyngeal residue	Summary grade created by identifying intersection between score on the variables; two variables – 'safety grade' and 'efficiency grade'	2 Volume, spatial, and patient response	Nominal scales which are modified by decision trees to produce to a 'grade' ranging from 0 (nil issues) to 4 (life-threatening); e.g. 
12 single variables (23), 2008	Poor bolus formation, prolonged oral transit, reduced velopharyngeal closure, delayed onset of swallow reflex, base of tongue / posterior pharyngeal wall weakness, reduced laryngeal elevation, reduced epiglottic inversion, reduced laryngeal vestibule closure, pharyngeal residue, cricopharyngeal muscle dysfunction, laryngeal penetration, aspiration	Nil summed score; nil subscales	12 Volume, temporal, and spatial	Dichotomous scale; abnormality 'present' or 'absent'
Single variable (Delay) (68), 2005	Timing swallow response	Nil summed score; nil subscales	1 Temporal	Raters completed three response options; time in seconds, a nominal scale indicating severity of delay ('mild', 'moderate' or 'severe') and dichotomous scale ('delayed' or 'not delayed')
Single Variables (Duration – bolus transit & Volume - residue) (67), 2006	Pharyngeal residue and bolus transit time	Nil summed score; nil subscales	2 Volume and temporal	Ordinal 3-point scale for valleculae and pyriform residue volume; e.g. '2: moderate residual with half the recess filled with residual post-swallow.' A continuous measure (time is seconds) used to evaluate transit time of bolus past anatomical landmarks

Measure; Reference, Year Published	Aspects evaluated by measure	Summed score / number of subscales[a]	Total number of items; domain of variables	Response Options
Unnamed - Single variable (Residue) (65), 2011	Pharyngeal residue	Number of structures affected is summed to create the variable's score	4 Spatial	Nominal scale, with associated scores ranging from 1 – 2; e.g. '1 = no residue +1 = valleculae residue + 2 = pyriform sinus +2 for posterior pharyngeal wall residue'
FEES and VFSS				
Penetration Aspiration Scale (PAS) (25), 1996	Location and volume of bolus in relation to airway and patient's response to penetration / aspiration	Nil summed score; nil subscales	1 Volume, spatial and patient response	Ordinal 8-point scale (1 – 8) with verbal descriptors; e.g. '2 – contrast enters the airway, remains above vocal folds; no residue'
University of California San Francisco (UCSF) Rating Form **(77)**, 2016	Amount and location of secretions and / or bolus residue across three anatomical categories (pharynx, larynx, trachea) which are divided into specific landmarks which were affected (e.g. laryngeal vestibule: upper 1/3). Utilised SEES procedure[b]	Nil summed score; nil subscales	7 (landmarks which may be affected). Volume and spatial	4-option nominal scale; absent, trace / minimal, moderate / maximal, unable to visualise. Raters referred to photographic exemplars
Single Variable (Volume - Residue) **(78)**, 2015	Presence or absence of pharyngeal residue	Nil summed score; nil subscales	1 Volume	Dichotomous scale; residue 'present' or 'absent' Pharyngeal residue defined as retention of greater than 15% of a given material in valleculae or pyriform sinuses

[a] Number of summed scores / subscales: summed score refers to all items or subscales results being considered collectively to produce an overall score / descriptor which describes the total performance or impact of the swallowing dys/function. Subscales refer to a subset of items being considered collectively to describe performance or designate score for a particular component of the swallow. Measures may have one summed score and multiple subscales.

[b] SEES: authors utilised Static Endoscopic Evaluation of Swallowing (SEES), a transoral rigid endoscopic procedure which produces images that are similar to FEES.

Table 7: Description of studies which report on multiple psychometric properties of measures

Measure; reference	Study on psychometrics	Study purpose	Study population, number (N)	Aetiologies, number (N)	Age (range, [R]) and / or Mean [M] years
Marionjoy 3- and 5-Point secretion severity scales					
Donzelli (69)	Donzelli (69)	Evaluate relationship between oropharyngeal sections and dysphagia diagnosis / diet recommendations; reduce the 5-point scale to the 3-point scale	Consecutive patients referred to otolaryngology / SLP departments (N = 100)	Neuromuscular impairment (N = 33), stroke (N = 30), dysphagia (N = 15), traumatic brain injury (N = 8), spinal cord / neck trauma (N = 7), neurosurgery (N = 4), anoxic encephalopathy (N = 3)	R = 10 – 81 M = 58.95
			Healthy controls (N = 4)	Nil history of dysphagia / head or neck abnormality	R = NR M = 46
Dysphagia Score					
Dziewas (74)	Dziewas (74)	Develop a scoring system for endoscopy which can guide dysphagia management (prescription of diet) and establish reliability data	Patients with first ever stroke (N = 100)	Stroke, within 24 hours of symptom onset	R = NR M = 70.5
P-Score					
Farneti (60)	Farneti (60)	Develop a scoring system for secretions / residue which is correlated to statistical data on aspiration	Acute, subacute, residential aged care in-patients and out-patients with and without aspiration referred to ENT (N = 520)	Stroke, traumatic brain injury, chronic cerebrovascular, post neurosurgery or maxilla-facial surgery, degenerative neurological disorders, elderly, children (N = NR)	R = NR M = 67.23
	Farneti (79)	Assess inter- and intra-rater reliability of the P-score	Consecutive out-patients (N = 23)	Globus (N = 1), cortical ictus sequelae (N = 5), reflux (N = 2), chronic obstructive pulmonary disease (COPD) (N = 2), dermamyositis (N = 1), laryngeal paralysis (N = 4), neurological degenerative (N = 2), corea major (N = 1), myasthenia (N = 1), head / neck surgery (N = 2), Sjogren's syndrome (N = 1), Wallemberg sequalae (N = 1)	R = 31 – 76 M = 58.56
BRACS					
Kaneoka (61)	Kaneoka (61)	Develop a scoring system to assess the amount / location of pharyngeal residue, patient response to residue and establish reliability and validity of the measure	In-patients and out-patients assessed for dysphagia (N = 51)	Head and neck cancer (N = 21), neurological diseases (N = 13), cardiovascular diseases (N = 7), respiratory diseases (N = 10), oesophageal diseases (N = 5), other (N = 7)	R = NR M = 61.4
Yale Pharyngeal Residue Severity Rating Scale					
Neubauer (75)	Neubauer (75)	Develop an image-based scoring system to assess the amount of valleculae and pyriform sinus residue	'13 images' of FEES from adults attending an urban hospital	NR	R = NR M = NR
Murray Secretion Severity Rating Scale					
Murray (76)	Murray (76)	Develop a scale to determine severity of secretions in hypopharynx to assist prediction of aspiration from instrumental assessment	Older hospitalised patients (N = 47)	COPD, diabetes mellitus or neurological pathology (N = NR)	R = 60 – 100 M = NR R = 60- 83

Measure; reference	Study on psychometrics	Study purpose	Study population, number (N)	Aetiologies, number (N)	Age (range, [R]) and / or Mean [M] years
			Older healthy non-hospitalised patients (N = 17)	NR	M = NR
			Younger, healthy participants (N = 5)	NR	R = 24 – 40 M = NR
	Marvin (44)	Determine if identification of penetration and aspiration differed between green-dyed and naturally white liquids	Hospitalised patients. Total (N = 40) Participants who completed trial of all textures (N = 19)	Cardiac surgery (N = 4), thoracic surgery (N = 4), head & neck surgery (N = 4), neurosurgery (N = 3), trauma (N = 3), septic shock (N = 3), organ transplant (N = 2), Guillain–Barre (N = 1), burns (N = 1), vascular surgery (N = 1)	♂ R = 28 – 86, M = 66 ♀ R = 42 – 78, M = 60
	Pluschinski (80)	Assess reliability and validity of the Murray Secretion Severity Rating Scale	Patients (N = 35)	NR	R = NR M = NR
PAS					
Rosenbek (25)	Butler (81)	Determine if PAS scores differ across bolus types (milks, water) and bolus size or delivery method	Healthy participants (N = 14)	No history of dysphagia, speech or voice disorders, pulmonary or neurologic diseases or structural disorders.	R = 69 - 85 M = 75
	Butler (82)	Determine reliability of the PAS as a function of clinician experience	35 swallow recordings	NR	R = NR M = NR
	Colodny (83)	Determine reliability of the PAS in FEES	79 swallow recordings	Stroke or other neurological disorders (70%), COPD and/or dementia (30%)	R= NR M = NR
	Daniels (67)	Develop a standard method of using VFSS to define dysphagia	Patients (N = 9 Healthy adults (N = 13)	Stroke Males with no history of neurological disease, COPD, head and neck cancer or dysphagia	R = 50 – 78 M = 62 R = 54 – 76 M = 64
	Hind (42)	Assess accuracy of PAS scoring made by hospital-based speech pathologists compared to unblinded expert judges	Patients who exhibited aspiration of thin liquids on VFSS (N = 669)	Parkinson's disease (49%), dementia (32%), both (19%)	R = 50 – 95 M= NR
	Kelly (84)	Determine if the type of examination (FEES vs VFSS) affects perception of penetration / aspiration	Patients referred for dysphagia assessment (N = 15)	Bilateral vocal-fold palsy (N = 1), suspected sarcoidosis (N = 1), cervical spine degeneration (N = 1), cerebral small vessel disease (N = 1), head and neck cancers (N = 5), none (N = 1) multiple sclerosis (N = 1), reflux (N = 2), systemic lupus erythematosus (N = 1), none (N = 1)	R = 22 – 78 M = 53.4
	McCullough (26)	Assess reliability of the PAS	Patients with stroke (N = 20)	Stroke within 6 weeks of VFSS	R = 40 - 96 M = 67.8

Measure; reference	Study on psychometrics	Study purpose	Study population, number (N)	Aetiologies, number (N)	Age (range, [R]) and / or Mean [M] years
	Omari (65)	Determine if bolus residue may be detected without use of VFSS	Patients with dysphagia (N = 23)	Stroke (N =7), cerebral palsy (N = 4), Parkinson's disease (N = 2), dementia (N = 2), neurosurgery (N = 1), cardiac disease (N = 1), motility disorders (N = 2) and unknown diagnoses (N =3)	R = 2 – 95 M = 55
			Healthy adults (N = 10)	No history of dysphagia or motility disorder	R = 24 – 47 M = 36.6
	Park (78)	Compare diagnostic efficacy between VFSS and endoscopist-directed FEES	Consecutive patients with suspected dysphagia (N = 50)	Stroke (N = 32), malignancy (N = 5), dementia (N = 4), deconditioning (N = 4), traumatic brain injury (N = 3), Parkinson's disease (N = 1), neuromuscular disease (N = 1)	R = 26 – 88 M = 67.8
	Rosenbek (25)	Define and describe use and development of the PAS and report reliability data	Patients with dysphagia (N = 15)	Stroke	R = NR M = NR
UCSF Rating Form					
Curtis (77)	Curtis (77)	Determine sensitivity and specificity of SEES compared to VFSS for assessing residue, penetration and aspiration	Consecutive patients presenting to UCSF voice and swallowing centre (N = 39)	Patients reporting dysphagia, globus, or chronic cough (N = NR)	R = NR M = NR
12 single variables					
12 single variables (23)	Frowen (23)	Compare the stability, reliability, and validity of three different types of measures used to analyse the VFSSs and determine if there is variability in psychometric properties across bolus textures	Patients within 3 months of treatment (N = 40)	Head and neck cancer (radiotherapy N = 10, chemotherapy N = 30)	R = 40 - 90 M = NR
MBSImp					
Martin-Harris (62)	Martin-Harris (62)	Establish the content, construct and external validity and inter- and intrarater reliability of the MBSImp	In and out-patients consecutively referred for swallow assessment (N = 300)	Pulmonary (23%), head and neck cancer (21%), neurology (16%), gastroenterology (12%), cardiothoracic (9%), general otolaryngology (5%), neurosurgery (3%), oncology (3%), general practice (3%), endocrine (2%), orthopaedics, trauma, general surgery, rheumatology, vascular, and unknown/unreported (<1% each)	R = NR M = NR
	Gullang (85)	Examine relationship between VFSS and manometry	Patients who completed both VFSS and manometry (N = 164)	Dysphagia (59%), choking sensation (15%), globus (11%), reflux (6%), aspiration pneumonia (4%), odynophagia (4%) and chronic cough (1%)	R = 21 - 94 M = 58
PDS					
Han (63)	Han (63)	Develop a quantitative functional dysphagia scale for stroke patients	Patients with symptoms of aspiration 3 days prior to VFSS (N = 103)	Stroke	R = 52 - 72 M = NR

Measure; reference	Study on psychometrics	Study purpose	Study population, number (N)	Aetiologies, number (N)	Age (range, [R]) and / or Mean [M] years
VDS					
Han (64)	Han (64)	Develop a measure to predict long-term prognosis of stroke patients with dysphagia	Patients within 72 hours of admission, repeated at 6 months post stroke (N = 83)	Stroke	R = 38 – 85 M = 62
	Kim (86)	Assess reliability of the VDS	Patients of rehabilitation centres (N = 100)	Stroke (N = 64), traumatic brain injury (N = 13), head and neck cancer (N = 12), brain tumours (N = 6) and other (N = 5)	R = NR M = 64.4
	Kim (87)	Determine the clinical applicability of the VDS to multiple aetiologies	Patients who underwent VFSS (N = 1, 995)	Stroke (N = 742), brain tumour (N = 199), neurodegenerative disease (N = 111), traumatic brain injury (N = 37), other brain disorders (N = 136), spinal cord injury (N = 37), neuromuscular junction disorder or myopathy (N = 52), peripheral neuropathy (N = 48), other (N = 279)	R = NR M = 58.7
DIGEST					
Hutcheson (66)	Hutcheson (66)	Explore feasibility and psychometrics of DIGEST	Patients post treatment (N = 100)	Head and neck cancers	R = 47 - 84 M = 61
Single variable (Delay)					
Karnell (68)	Karnell (68)	Assess reliability of clinician's judgements of swallow delay compared to temporal measures	Patients with dysphagia without structural abnormalities or absent swallow (N = 20)	Throat irritation (N = 1), reflux (N = 1), Hashimoto's disease (N = 1), brain cancer (N = 1), sarcoidosis (N = 1), chronic cough / throat irritation (N = 3), globus (N = 1), right hemiparesis (N =1), stroke (N = 4), multiple sclerosis (N = 1), dental issues (N = 1), oesophageal stenosis (N = 1), pneumonia (N = 2), coughing while eating / drinking (N = 1)	R = 29.7 - 83 M = 61.6
Single Variables (Duration – bolus transit & Volume - residue)					
Daniels (67)	Daniels (67)	See Daniels, under PAS	**RE**	**RE**	**RE**
Single variable (Residue)					
Omari (65)	Omari (65)	See Omari, under PAS	**RE**	**RE**	**RE**
Single Variable (Volume - residue)					
Park (78)	Park (78)	See Park, under PAS	**RE**	**RE**	**RE**

Note: NR = not reported; RE = reported elsewhere

Table 8: Overview of the methodological quality assessment results using the COSMIN checklist: studies reporting on psychometric properties of VFSS and FEES measures

Measure & Author(s)	Internal Consistency[a]	Reliability	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
Marionjoy 3-Point secretion severity scale						
Donzelli et al. (69): total scale	N/A	NR	NR	Fair (50.0%)	NR	NR
Penetration	N/A	NR	NR	NR	NR	Fair (30.4%)
Aspiration	N/A	NR	NR	NR	NR	Fair (30.4%)
Diet Outcomes	N/A	NR	NR	NR	NR	Fair (29.4%)
Marionjoy 5-Point secretion severity scale						
Donzelli et al. (69): total scale	N/A	Fair (27.3%)	NR	Good (71.4%)	NR	NR
Penetration	N/A	NR	NR	NR	NR	Fair (34.8%)
Aspiration	N/A	NR	NR	NR	NR	Fair (34.8%)
Diet Outcomes	N/A	NR	NR	NR	NR	Fair (47.1%)
Tracheostomy Status	N/A	NR	NR	NR	NR	Fair (47.1%)
Dysphagia Score[b]						
Dziewas et al. (74)	NR	Fair (27.2%)	NR	Fair (42.9%)	NR	NR
P-Score[c]						
Farneti (60)	NR	NR	NR	Good (57.1%)	NR	NR
Farneti et al. (79)	NR	Fair (42.43)	NR	NR	NR	NR
BRACS						
Kaneoka et al. (61)	Good (71.4%)	Good (57.6%)	NR	Good (71.4%)	Good (58.3%)	Fair (39.1%)
Yale Pharyngeal Residue Severity Rating Scale						
Neubauer et al. (75)	NR	Excellent (81.8%)	NR	Fair (35.7%)	NR	Fair (43.5%)
Murray Secretion Severity Scale						
Murray et al. (76)	N/A	Fair (26.7%)	NR	Excellent (78.6%)	NR	NR
Pluschinski et al. (80)	N/A	Good (54.5%)	NR	NR	NR	Fair (30.4%)
Marvin et al. (44)	N/A	Fair (31.25)	NR	NR	NR	NR
Single Variable (Volume - residue)[b]						

Measure & Author(s)		Internal Consistency[a]	Reliability	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
Park et al. (78)	Pharyngeal residue - viscous food	N/A	NR	NR	NR	NR	Poor (21.7%)
	Pharyngeal residue – overall	N/A	NR	NR	NR	NR	Poor (21.7%)
	Pharyngeal residue – liquids	N/A	NR	NR	NR	NR	Poor (21.7%)
Standardised Grading Forms							
Curtis et al. (77)		NR	Fair (33.33)[d]	NR	NR	NR	Poor (13.04)
Single Variables (Duration & Volume - residue)							
Daniels et al. (67)	Bolus duration (s)	NR	Poor (24.1%)	NR	Fair (35.7%)	NR	NR
	Bolus clearance	NR	Poor (15.1%)	NR	Fair (28.6%)	NR	NR
MBSImp							
Martin-Harris et al. (62)		NR	NR	NR	Good (64.3%)	Excellent (83.3%)	Good (65.2%)
Gullang et al. (85)		NR	NR	NR	NR	NR	Poor (21.1%)
VDS							
Han et al. (64)		NR	NR	NR	Fair (50.0%)	NR	NR
Kim et al. (87)		NR	NR	NR	NR	NR	Fair (47.8%)
Kim et al. (86)		NR	Fair (31.58)	NR	NR	NR	NR
FDS							
Han et al. (63)		NR	Fair (44.8%)	NR	NR	NR	Fair (30.4%)
DIGEST							
Hutcheson et al. (66)		NR	Good (63.3%)	NR	Excellent (100%)	NR	Fair (43.5%)
PAS – FEES							
Butler et al. (81)		N/A	Excellent (81.82%)	NR	NR	NR	NR
Butler et al. (82)		N/A	Fair (42.42%)	NR	NR	NR	NR
Colodny (83)		N/A	Good (54.55%)	NR	NR	NR	NR
Kelly et al. (84)		N/A	Good (60.61%)	NR	NR	NR	NR
Park et al. (78)		N/A	NR	NR	NR	NR	Poor (21.7%)
PAS – VFSS							
Daniels et al. (67)		N/A	Poor (15.1%)	NR	NR	NR	NR

Measure & Author(s)		Internal Consistency[a]	Reliability	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
Hind et al. (42)		N/A	Fair (36.84%)	NR	NR	NR	NR
Kelly et al. (84)		N/A	Good (60.61%)	NR	NR	NR	NR
McCullough et al. (26)		N/A	Fair (33.33%)	NR	NR	NR	NR
Omari et al. (65)		N/A	NR	NR	NR	NR	Good (52.2%)
Park et al. (78)		N/A	NR	NR	NR	NR	Poor (21.7%)
Rosenbek et al. (25)		N/A	Good (66.66%)	NR	Good (57.14%)	NR	NR
Single variable (Timing - delay)							
Karnell et al. (68)	Latency (s)	N/A	Good (64.0%)	NR	NR	NR	Fair (39.1%)
	Dichotomous options	N/A	Good (63.6%)	NR	NR	NR	Fair (39.1%)
	Severity	N/A	Good (63.6%)	NR	NR	NR	NR
Single variable (residue)							
Omari et al. (65)		NR	NR	NR	NR	NR	Fair (47.8%)
12 Single Variables (Spatial, Timing and Volume)							
Frowen et al. (23)	Semi-solids	NR	Good (57.6%)	NR	NR	NR	Good (60.1%)
	Liquids	NR	Good (57.6%)	NR	NR	NR	Good (60.1%)

Notes:
The quality of the studies that evaluated the psychometric properties of each measure was evaluated according to the COSMIN rating per item: four-point scale was used (1 = Poor, 2 = Fair, 3 = Good, 4 = Excellent). The overall methodological quality per study was presented as percentage of rating (Poor = 0–25.0%, Fair = 25.1%- 50.0%, Good = 50.1%-75.0%, Excellent = 75.1%-100.0%)
NR: not reported
N/A: not applicable

[a]Measures which utilised only one item were unable to be assessed for internal consistency; this property is marked not applicable (N/A) for these studies

[b]Measure likely not developed in English, although study published in English. Attempted to contact author; no information available on translation process.

[c]Measure developed in Italian, published in English. Authors report the P-score utilises only five anatomical terms (e.g. vallecula marginal zone, pyriform sinus), three volume terms (coating, minimum, maximum) and 3 quantity terms (< 2, 2 > < 5, >5) all of which have direct equivalents in English. COSMIN translation score: 27.77% (Fair)

[d]Score pertains reliability for SEES only

Table 9: Quality of psychometric properties per study based on the criteria by Terwee et al. (36) and Schellingerhout (38)

Measure & Author(s)	Internal Consistency	Reliability Inter:	Reliability Intra:	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
Marionjoy 3-Point secretion severity scale							
Donzelli et al. (69)	N/A	NR	NR	NR	?	NR	?
Marionjoy 5-Point secretion severity scale							
Donzelli et al. (69)	N/A	+	NR	NR	?	NR	?
Dysphagia Score							
Dziewas et al. (74)	NR	+	NR	NR	+	NR	NR
P-Score							
Farneti (60)	NR	NR	NR	NR	?	NR	NR
Farneti et al. (79)	NR	+	+	NR	NR	NR	NR
BRACS							
Kaneoka et al. (61)	?	+	+	NR	?	+	?
Yale Pharyngeal Residue Severity Rating Scale							
Neubauer et al. (75)	NR	+	+	NR	?	NR	?
Murray Secretion Severity Scale							
Murray et al. (76)	N/A	?	NR	NR	?	NR	NR
Pluschinski et al. (80)	N/A	?	?	NR	NR	NR	?
Marvin et al. (44)	N/A	NR	?	NR	NR	NR	NR
Standardised Grading Forms							
Curtis et al. (77)	NR	?	?	NR	NR	NR	NE
MBSImp							
Martin-Harris et al. (62)	NR	NR	NR	NR	?	?	?
Gullang et al. (85)	NR	NR	NR	NR	NR	NR	NE
VDS							
Han et al. (64)	NR	NR	NR	NR	+	NR	NR
Kim et al. (87)		NR	NR	NR	NR	NR	?
Kim et al. (86)		-	-	NR	NR	NR	NR
FDS							

Measure & Author(s)		Internal Consistency	Reliability Inter:	Reliability Intra:	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
Han et al. (63)		NR	+	NR	NR	NR	NR	?
DIGEST								
Hutcheson et al. (66)		NR	-	+	NR	+	NR	?
PAS – FEES								
Butler et al. (81)		N/A	-	NR	NR	NR	NR	NR
Butler et al. (82)		N/A	+	+	NR	NR	NR	NR
Colodny (83)		N/A	±	+	NR	NR	NR	NR
Kelly et al. (84)		N/A	±	±	NR	NR	NR	NR
Park et al. (78)		N/A	NR	NR	NR	NR	NR	NE
PAS – VFSS								
Daniels et al. (67)		N/A	NE		NR	NR	NR	NR
Hind		N/A	+	NR	NR	NR	NR	NR
Kelly et al. (84)		N/A	±	±	NR	NR	NR	NR
McCollough et al. (26)		N/A	±	?	NR	NR	NR	NR
Omari et al. (65)		N/A	NR	NR	NR	NR	NR	?
Park et al. (78)		N/A	NR	NR	NR	NR	NR	NE
Rosenbek et al. (25)		N/A	±	±	NR	?	NR	NR
Single Variables (Temporal and Volume - residue)								
Daniels et al. (67)	Bolus Duration	NR	NE	NE	NR	?	NR	NR
	Residue	NR	NE	NE	NR	?	NR	NR
Single variable (Volume - residue)								
Omari et al. (65)		NR	NR	NR	NR	NR	NR	?
Single Variable (Volume - residue)								
Park et al. (78)		N/A	NR	NR	NR	NR	NR	NE
Single variable (delay)								
Karnell et al. (68)	Latency (s)	N/A	?	?	NR	NR	NR	?
	Dichotomous options	N/A	±	±	NR	NR	NR	?
	Severity	N/A	-	-	NR	NR	NR	NR

Measure & Author(s)		Internal Consistency	Reliability Inter:	Reliability Intra:	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
12 Single Variables (Spatial, Timing and Volume – dichotomous options)								
Frowen et al. (23)	Semi-Solids	NR	?	?	NR	NR	NR	?
	Liquids	NR	?	?	NR	NR	NR	?

Quality criteria (38): ? = positive rating; ? = indeterminate rating; - = negative rating; ± = conflicting data; NR = not reported; NE = not evaluated (study of poor methodological quality according to COSMIN rating—data are excluded from further analyses)

Table 10: Overall quality score of assessments for each psychometric property based on levels of evidence by Schellingerhout et al. (38)

Measure; reference	Internal Consistency	Reliability	Measurement Error	Content Validity	Structural Validity	Hypothesis testing
FEES						
Marionjoy 3-Point secretion severity scale Donzelli et al. (69)	N/A	NR	NR	Indeterminate	NR	Indeterminate
Marionjoy 5-Point secretion severity scale Donzelli et al. (69)	N/A	Limited (positive)	NR	Indeterminate	NR	Indeterminate
Dysphagia Score Dziewas et al. (74)	NR	Limited (positive)	NR	Limited (positive)	NR	NR
P-Score Farneti (60)	NR	Limited (positive)	NR	Indeterminate	NR	NR
BRACS Kaneoka et al. (61)	Indeterminate	Moderate (positive)	NR	Indeterminate	Moderate (positive)	Indeterminate
Yale Pharyngeal Residue Severity Rating Scale Neubauer et al. (75)	NR	Strong (positive)	NR	Indeterminate	NR	Indeterminate
Murray Secretion Severity Scale Murray (76)	N/A	Indeterminate	NR	Indeterminate	NR	Indeterminate
Standardised Grading Forms Curtis (77)	NR	Indeterminate	NR	NR	NR	NE
PAS Rosenbek et al. (25)	N/A	Conflicting	NR	NR	NR	NE
VFSS						
MBSImp Martin-Harris et al. (62)	NR	NR	NR	Indeterminate	Indeterminate	Indeterminate
VDS Han et al. (64)	NR	Limited (negative)	NR	Limited (positive)	NR	Indeterminate
FDS Han et al. (63)	NR	Limited (positive)	NR	NR	NR	Indeterminate
DIGEST Hutcheson et al. (66)	NR	Conflicting	NR	Strong (positive)	NR	Indeterminate
PAS Rosenbek et al. (25)	N/A	Conflicting	NR	Indeterminate	NR	Indeterminate
Single Variables (Temporal and Volume) Daniels et al. (67)	NR	NE	NR	Indeterminate	NR	NR
Single variable (Volume - residue) Omari (65)	NR	NR	NR	NR	NR	Indeterminate
Single variable (Temporal) Karnell (68)	N/A	Conflicting	NR	NR	NR	Indeterminate
12 single variables Frowen (23)	NR	Indeterminate	NR	NR	NR	Indeterminate

Notes:
Levels of Evidence: Strong evidence positive/negative result = Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality; Moderate evidence positive/negative result = Consistent findings in multiples studies of fair methodological quality OR in one study of good methodological quality; Limited evidence positive/negative = One study of fair methodological quality; Conflicting findings; Indeterminate = only indeterminate measurement property ratings (i.e., score = ? in Table 3); NR = Not reported; Not Evaluated = studies of poor methodological quality according to COSMIN excluded from further analyses.

Supplementary Table 1

PRISMA 2009 Checklist			
Section/topic	**#**	**Checklist item**	**Reported on page #**
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	7
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	8
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	8-9, table 1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	9
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	9
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	9
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	11

Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	9-11, Tables 2-4
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9-11, Tables 2-4
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	11
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating, which were pre-specified.	12-15
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Figure 2
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Tables 5-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Tables 5-7
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	33-39
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Tables 8 - 10
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	16-22
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	22
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Figure 1: Methodological quality and psychometric properties analysis process

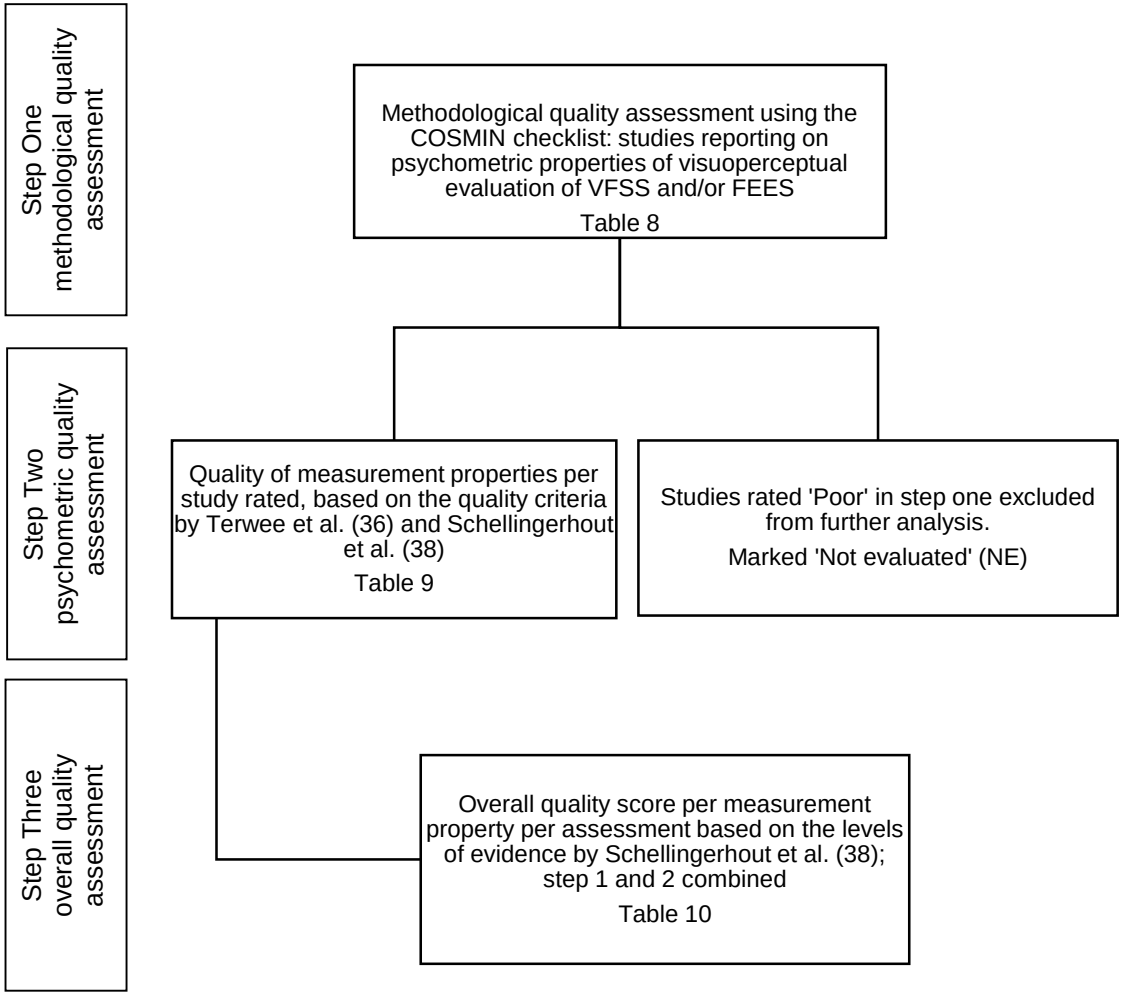


Figure 2: Flow diagram of reviewing process according to PRISMA (1)