Dysphagia (2019) 34:2–33
https://doi.org/10.1007/s00455-018-9918-3

REVIEW ARTICLE

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

Katina Swan[1] · Reinie Cordier[1] · Ted Brown[2] · Renée Speyer[1,3,4]

Received: 29 November 2017 / Accepted: 4 June 2018 / Published online: 17 July 2018

Abstract

Fibreoptic Endoscopic Evaluation of Swallowing (FEES) and Videofluoroscopic Swallow Studies (VFSS) are instrumental assessments of dysphagia which provide 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 that 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. This review assessed the quality of the psychometric properties of visuoperceptual measures of FEES and VFSS. 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 the methodical quality of studies. The measures' overall quality was then assessed by combining COSMIN ratings with quality criteria. Forty-five studies, reporting on 39 measures, met the inclusion criteria for this review. Data about the measures' psychometric properties were very limited. Twenty-one measures had information available about reliability only, while 18 had information on up to five of the possible nine psychometric properties 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, issues with design, statistical analyses, and reporting practices of extant studies. 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, which utilises robust methodological design and reporting, is needed to examine the psychometric properties of measures for FEES and VFSS.

Keywords 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 [1–4]. 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% [5]. In a 1-year period, approximately 9.5 million adults in the United States reported a swallowing problem [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

PROSPERO Registration No: CRD42017060032.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00455-018-9918-3) contains supplementary material, which is available to authorised users.

✉ Katina Swan
katina.swan@postgrad.curtin.edu.au

Extended author information available on the last page of the article

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, especially at special social events where eating 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 at the bedside, where clinicians gather patient history and concerns, and use non-invasive testing to assess nerve and muscle function to establish the pattern of impairment [16]. However, these assessments have limitations in terms of the breadth and accuracy of information they are able to provide. Since swallowing is an internal process, 'bedside' or clinical assessment does 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 Endoscopic 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 ten 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]. 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 management of secretions, food and fluid boluses; the ability to perform swallow manoeuvres; the presence of any structural abnormalities; and the impact of the dysphagia on swallowing safety.

This interpretation of recordings produced by VFSS and FEES typically involves the dysphagia clinician viewing the recordings several times and making subjective judgements, which are 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 'objective' assessments, their interpretation is subjective, as 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 VFSS 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 with reference to anatomy, or the size and scale of a clinically relevant indicator), volume (amount of bolus or secretions involved), and patient response variables (such as coughing or 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 penetration and aspiration visualised in VFSS or 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; further, diagnosis and decisions about patient care may be compromised.

In an initiative to evaluate the quality of the psychometric properties of measures commonly used to analyse VFSS, McCullough et al. [26] reviewed the following: 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 VFSS. 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 VFSS measures were appropriate for use in clinical and research settings. These studies, while representing a promising start, 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 on international consensus, for the assessment of the quality of studies which examine 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 issues 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) to assess both the methodological quality of studies reporting on such measures and the quality of the psychometric properties of these measures, and (3) to synthesise this information to indicate current state of knowledge about the psychometric soundness of visuoperceptual measures of VFSS and FEES. This systematic review focuses on measures that were published in English and which assess visuoperceptual aspects of recordings of 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 [29]. See Supplementary Table 2 for the completed PRISMA checklist for this review.

Eligibility Criteria

Studies eligible for inclusion were published 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 of any age, visuoperceptual measure/s which analysed data from VFSS or FEES, report on the 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 as the authors aimed to focus on measures most likely to be used in clinical practice. Although there are several software programmes available to assist analysis, which offer a more objective interpretation of VFSS and FEES [30], they are often limited in terms of clinical use due to the considerable time required to use them [20]. VFSS and FEES clinics typically see multiple patients consecutively, due to limited availability of the equipment and various clinical staff required [31], making routine use of potentially time-consuming software impractical.

Each instrument was evaluated for reliability and validity according to the COSMIN taxonomy of measurement properties and definitions for health-related patient-reported outcomes [32]. 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, 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 Speyer 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. Supplementary Table 1 lists search terms used across all databases. References of articles accepted to the review were hand searched for additional suitable studies. Refer to Supplementary Table 1, for list of search terms.

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 was 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 were discussed by reviewers until consensus was met. Author Speyer 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, with a 40% random selection evaluated by an independent second reviewer. Full texts were likewise excluded if they did not meet criteria (see Fig. 2). There was 100% consensus between reviewers.

Data Collection Process and Data Extraction

Measures fell into two categories: (1) measures with studies which provided information on inter- and intra-rater reliability only, and (2) measures with studies which reported on multiple psychometric properties or properties other than inter- and intra-rater reliability. 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. inter-rater reliability and content validity) or properties other than inter- and intra-rater reliability 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 were guided by the Cochrane Handbook for Systematic Reviews [33, Sect. 7.3a] and the Systematic Reviews Centre for Reviews and Dissemination [34].

Methodological Quality

The methodological quality of the included studies was assessed using the COSMIN taxonomy of measurement properties and definitions for health-related patient-reported outcomes [32, 35]. 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, hypothesis testing, cross-cultural validity, criterion validity, and responsiveness [32]. Refer to Table 1 for the definitions of all psychometric properties as defined by the COSMIN statement [35]. Criterion validity was not evaluated due to the absence of a 'gold-standard' measure for FEES and VFSS. Responsiveness was beyond the scope of this review; further, 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; however, where the original measure was developed in a language other than English, quality of translation process was assessed.

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 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 checklist items only have an option to rate the item as 'Good' or 'Excellent', the total score for each psychometric property was calculated

Table 1 COSMIN definitions of domains, psychometric properties, and aspects of psychometric properties for health-related patient-reported outcomes adapted from Mokkink et al. [95]

Psychometric property	Domain
Definition	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 among items
Reliability	The proportion of total variance in the measurements due to "true" differences among 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

[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

using the formula detailed below, to accurately capture the quality scores [32]:

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

After the methodological quality of studies was assessed, those psychometrics properties 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 these psychometric properties. Studies that received a 'Poor' COSMIN rating were excluded from further analysis, as results arising from studies using flawed methodology were considered unreliable. Table 2 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 methodological quality scores measured by COSMIN and the psychometric quality ratings as measured by Terwee et al. [36] and Schellingerhout [38]; refer to Table 3. This is consistent with methodology utilised in previous psychometric reviews [39, 40]. Refer to Fig. 1 flow chart for overview of the analysis process.

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 criterion validity, 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 by (a) the use of the COSMIN checklist, an internationally recognised standard for rating study quality of psychometric studies; (b) the use of pre-established

Table 2 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
	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 hypothesis testing
	NE	Not evaluated
Internal consistency	+	Factor analyses performed on adequate sample size (7 *# items consistency and $\geq$ 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 $\geq$ 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 3 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: '?') and therefore also indeterminate level of evidence for the overall quality of that measurement property

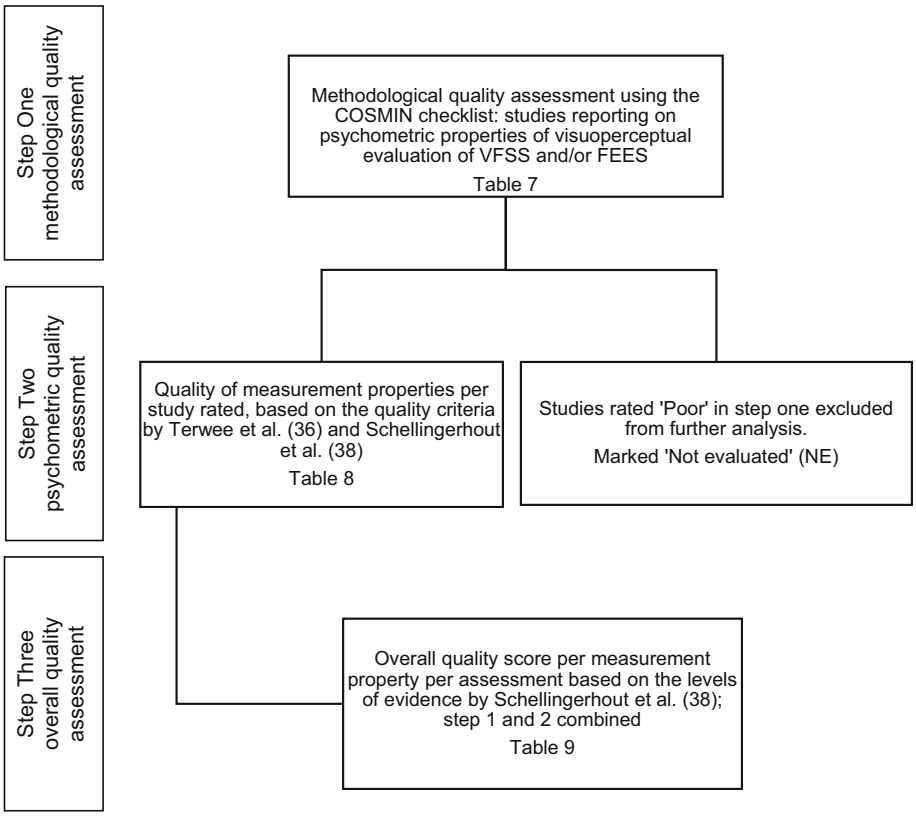

Fig. 1 Methodological quality and psychometric properties analysis process

criteria of quality of psychometric properties; and (c) with the rating of psychometric properties of a 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] quality criteria ratings, reviewers convened until consensus was achieved. Author Speyer 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 4, 5, and 6 display the synthesised data collected from each measure and article reporting on psychometric properties.

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,

Table 4 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**	**Variables evaluated by measure**	**Total number of items[a]; Domain of variables**	**Response options**	**COSMIN quality score**	**Quality of psychometric properties: inter / intra - rater reliability**	**Overall quality score**
			FEES				
Unnamed Marvin et al. (41), 2016	Marvin et al. (41)	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 (42.4%)	Inter: NR Intra: Using green bolus : + Using white bolus : -	Limited (positive) Limited (negative)
Unnamed Pilz et al. (42), 2016[c]	Pilz et al. (42)	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'	Good (73.7%) Excellent (78.8%)	Inter : ± Intra : +	Conflicting
Unnamed Rodiriguez et al. (43), 2007	Rodiriguez et al. (43)	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 (36.8%) Fair (48.5%)	Inter :? Intra : +	Indeterminate
Unnamed[c] Susa et al. (44), 2015	Susa et al. (44)	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 (36.8%) Good (51.5%)	Inter : + Intra : +	Limited (positive)
Physiological and swallowing evaluation form[c] Tohara et al. (45), 2010	Tohara et al. (45)	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 (63.6%) Good (72.7%)	Inter: - Intra: ±	Conflicting
Unnamed Warnecke et al. (46), 2016	Warnecke et al. (46)	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'	Fair (31.6%) Good (57.6%)	Inter: + Intra : +	Limited (positive)
			VFSS				
Unnamed Bryant et al. (47), 2012	Bryant et al. (47)	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') Note: 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 (52.6%)	Inter : - Intra : NR	Moderate (negative)
Bethlehem Assessment Scale (BAS) Scott (59), 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'	(Fair) 26.3% Good (57.6%)	Inter : + Intra : +	Limited (positive)
Unnamed Gibson et al. (48) 1995	Gibson et al. (48)	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 (45.5%) Fair (50.0%)	Inter : ± Intra : ±	Conflicting

Table 4 continued

					Reliability[b]		
Measure; Reference, Year published	**Study on psychometrics**	**Variables evaluated by measure**	**Total number of items[a]; Domain of variables**	**Response options**	**COSMIN quality score**	**Quality of psychometric properties: inter / intra - rater reliability**	**Overall quality score**
			FEES				
Temporal and Physiologic Features of Infant Swallows Gosa et al. (49), 2015	Gosa et al. (49)	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 dichotomous options; e.g. 'epiglottic tilting: yes / no'	Poor (18.2%) Fair (47.2%)	Inter : NE Intra : ?	Indeterminate
Unnamed Hind et al. (50), 2009	Hind et al. (50)	Presence or absence of aspiration	1 Spatial	Dichotomous options of presence / absence of aspiration	Good (52.6%)	Inter: + Intra : NR	Moderate (positive)
'Objective measures' based on norms from Leonard et al. (96) Lee et al. (51)	Lee et al. (51)	Hyoid elevation, pharyngeal area, pharyngeal constriction ratio and pharyngo-oesophageal segment opening	4 Spatial	Dichotomous options of normal / abnormal	Good (55.0%) Good (60.6%)	Inter : - Intra : ±	Conflicting
Unnamed Mann et al. (53), 2000	Mann et al. (53)	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 (68.4%)	Inter: Diagnosis of dysphagia: + Inter: Diagnosis of aspiration: - Intra : NR	Moderate (positive) Moderate (negative)
Unnamed Miles (55), 2016	Miles (55)	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 (36.8%)	Inter: ± Intra : NR	Conflicting
Unnamed McCullough et al. (54),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	Good (60.0%) Fair (48.5%)	Inter: - Intra: ±	Conflicting
'VFSS objective measures', adapted from Leonard and Kendall (52),1997	Nordin et al. (62)	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	Poor (20.0%) Good (51.7%)	Inter: NE Intra : +	Moderate (positive)
Unnamed Power et al. (56), 2009	Power et al. (56)	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%)	Inter: NR Intra: +	Moderate (positive)
Bolus residue scale[c] Rommel et al. (58), 2015	Rommel et al. (58)	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 (26.3) Fair (33.3%)	Inter: + Intra: +	Limited (positive)
Modified Charing Cross Hospital Dysphagia Profile Price et al, 1987 (57)	Scott et al. (63)	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 (15.8%)	Inter: NE Intra : NR	NE
Unnamed[c] Stoeckli et al. (60), 2003	Stoeckli et al. (60)	Lip closure, soft palate / posterior tongue seal, bolus transport / lingual motion, delayed initiation, soft palate elevation, tongue base retraction, laryngeal elevation, laryngeal closure, anterior hyolaryngeal excursion, pharyngeal contraction, upper oesophageal sphincter opening / closure, penetration / aspiration, residue	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 Residue location: floor of mouth, base of tongue, valleculae, pharyngeal wall, aryepiglottic folds, pyriform sinuses'	Fair (47.4%)	Inter: - Intra : NR	Limited (negative)

Table 4 continued

FEES and VFSS							
Pharyngeal Residue Severity Scale Kelly et al. (61), 2006	Kelly et al. (61)	Volume of pharyngeal residue	1 Volume	Nominal scale reporting volume of residue: 'none', 'coating', 'mild', 'moderate' or 'severe'	Good (70.0%) Good (63.6%)	Inter : - Intra : +	Conflicting

[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* attempted to contact all authors; no information available on translation process, with the exception of Pilz et al. [42]

Pilz et al. [42] reported the measure was originally created in Dutch, and then subsequently translated to English using a professional translator. Translation process score according to COSMIN: 33.33% (Fair)

PubMed = 1429. Abstract duplicates totalled 293. Duplicates were removed and 1797 abstracts were screened for inclusion in the review, with 1581 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 reasons for exclusion of abstracts and full texts.

Included Measures

Due to the limited information available about their psychometric properties, measures where information is available solely on inter- and intra-rater reliability are presented separately (Table 4) from the measures with information about multiple psychometric properties or properties other than inter- and intra-rater reliability (Tables 5, 6). These were collated separately, as measures with known psychometric properties for reliability and validity are likely to be more relevant to the clinician or researcher.

Table 4 synthesises the characteristics of these 21 inter- and intra-rater reliability only measures. Six measures analysed FEES recordings only [41–46]; 14 measures were for VFSS recordings [47–60] and one analysed both FEES and VFSS recordings [61], i.e. 7 measures of FEES and 15 measures of VFSS. FEES measures most commonly included variables related to aspiration, penetration, secretions, and residue (5 of 7 measures; [41, 42, 45, 46]), while VFSS measures most commonly had variables related to pharyngeal residue (10 of 15 measures; [47–49, 54, 57–61]), aspiration (8 of 15 measures; [47–50, 53, 57, 59]), timing of swallow initiation (7 of 15 measures; [47, 48, 53, 54, 56, 59, 60]), pharyngeal phase duration (7 of 15 measures; [47–49, 52, 54, 56, 59]), and oral phase duration (7 of 15 measures; [47–49, 53, 54, 56, 59]). 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 [47, 55]. None of the measures utilised summed scores or subscales; all were composed of one or more single variables. With the exception of Gosa et al. [49], all studies recruited adult populations only. Overall, the majority of measures (16 of 21; [41–50, 53, 55, 56, 58, 60, 61]) 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) the 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) [41, 43–45, 47–49, 53, 60, 61] and ordinal scales with associated descriptors at each level (n = 9; e.g. secretion colour: clear, white, brown, yellow, or bloody' or '0 = no pooling, 1 = filling of < 50% of the valleculae, 2 = filling of > 50% of valleculae) [42, 45–47, 49, 57–60]. Other options included dichotomous scales (n = 6; e.g. aspiration present: yes/no) [43, 49–51, 54, 55] and open-ended response options [48, 54, 56, 62], where raters recorded their judgements of continuous variables, such as time taken to complete a swallow phase (n = 4). The number of items utilised in FEES measures ranged from one to 16

Table 5 Description of characteristics of MEASURES with data on multiple psychometric properties, or a property other than reliability

Measure; reference, year published	Variables evaluated by measure	Summed score/ number of subscales[a]	Total number of items; domain of variables	Response options
FEES				
Marianjoy 3-point secretion severity scale Donzelli et al., 2003 [64]	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'
Marianjoy 5-point secretion severity scale Donzelli et al., 2003 [64]	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 Dziewas et al., 2008 [65]	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) Farneti, 2008 [77]	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) Kaneoka et al., 2013 [66]	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
Murray secretion scale Murray et al.,1996 [67]	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'

Table 5 continued

Measure; reference, year published	Variables evaluated by measure	Summed score/ number of subscales[a]	Total number of items; domain of variables	Response options
Yale pharyngeal residue severity rating scale Neubauer et al., 2015 [68]	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'
VFSS				
Functional dysphagia scale (FDS) Han et al., 2001 [70]	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- fluoroscopic dsyphagia scale (VDS) Han et al., 2008 [71]	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) Hutcheson et al.., 2017 [27]	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. Nominal scales which are modified by decision trees to produce to a 'grade' ranging from 0 (nil issues) to 4 (life - threatening); e .g. Maximum percentage of pharyngeal residue: >90% near complete residue' Pattern of residue: Any (but not all) bolus types / All bolous types Efficiency Grade = Grade 3 / Grade 4

Table 5 continued

Measure; reference, year published	Variables evaluated by measure	Summed score/ number of subscales[a]	Total number of items; domain of variables	Response options
MBS impairment tool (MBSImp) Martin-Harris et al., 2008 [73]	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 stripping wave, pharyngeal contraction, cricopharyngeal opening, tongue base retraction, pharyngeal residue and oesophageal clearance	Nil summed score; seventeen 'components' which are individually rated for each bolus texture	17; volume, temporal, spatial and patient response	3–5-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' 'Overall impression' score per swallow component also applied, which derives from scores across multiple bolus presentations
Two variables (duration–bolus transit & volume–residue) Daniels at al., 2006 [69]	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
Twelve variables Frowen et al., 2008 [23]	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) Karnell and Rogus, 2005 [21]	Timing of 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 variable (residue–location) Omari et al., 2011 [74]	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'

Table 5 continued

Measure; reference, year published	Variables evaluated by measure	Summed score/ number of subscales[a]	Total number of items; domain of variables	Response options
FEES and VFSS				
University of California San Francisco (UCSF) Standardised grading form Curtis et al., 2016 [75]	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 (residue–volume) Park et al., 2015 [76]	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
Penetration aspiration scale (PAS) Rosenbek et al., 1996 [25]	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'

[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 dysfunction. 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

(mean = 4.4). VFSS measures used a greater range, from one to 23 (mean = 8.3). In total, 14 measures used less items than the mean for their respective instrumental assessment; of these, eight received overall positive scores [41, 44, 46, 50, 53, 56, 58, 62]. In contrast, seven measures [26, 45, 47, 49, 60, 63] used more items than the mean and only one received a positive score for reliability overall [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) [41, 53]; both scored positive for reliability overall in only one protocol or diagnosis (green bolus and dysphagia, respectively).

Table 5 describes the characteristics of the 18 measures with known multiple psychometric properties or properties other than reliability only. Seven measures pertained to FEES only [23, 64–68] and eight measures analysed VFSS only [21, 23, 69–74]; three measures pertained to both FEES and VFSS [25, 75, 76]. 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) [64–68, 75–77]. Two measures assessed penetration/aspiration [25, 75], with patient response to airway invasion assessed by three measures [25, 65, 67]. Two measures utilised a summed score or subscales to formulate overall ratings: P-Score [77] and the BRACS [66]. The remainder did not use summed scores/subscales. Among measures of VFSS, the most commonly analysed variables were pharyngeal residue (n = 9) [23, 69–76], swallow reflex initiation (n = 5) [21, 23, 70, 71, 73], penetration/aspiration (n = 5) [23, 25, 70–72], laryngeal/hyoid elevation (n = 4) [23, 70, 71, 73], bolus formation/control (n = 4) [23, 70, 71, 73], epiglottic movement (n = 4) [23, 70, 71, 73], oral transit duration (n = 3) [23, 70, 71], pharyngeal transit duration (n = 3) [69–71], and lip closure (n = 3) [70, 71, 73]. Similar to the measures that reported on inter-/intra-rater reliability only (Table 4), oesophagus function was the most rarely included variable, with only one measure including analysis of the oesophageal phase swallow [73]. Consistent with FEES measures, VFSS

Table 6 Description of STUDIES which report on multiple psychometric properties of measures, or a property other than reliability

Measure; reference	Study on psychometrics	Study purpose	Study population, number (*N*)	Aetiologies, number (*N*)	Age (range, [*R*]) and/or Mean [*M*] years
FEES					
Marianjoy 3- and 5-point secretion severity scales					
Donzelli et al., 2003 [64]	Donzelli et al. [64]	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 et al., 2008 [65]	Dziewas et al. [65]	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 h of symptom onset	*R* = NR *M* = 70.5
P- score					
Farneti, 2008 [77]	Farneti [77]	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 et al. [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 et al., 2013 [66]	Kaneoka et al. [66]	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

Table 6 continued

Measure; reference	Study on psychometrics	Study purpose	Study population, number (*N*)	Aetiologies, number (*N*)	Age (range, [*R*]) and/or Mean [*M*] years
Murray secretion scale					
Murray et al., 1996 [67]	Murray et al. [67]	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
			Older healthy non-hospitalised patients (*N* = 17)	NR	*R* = 60–83 *M* = NR
			Younger, healthy participants (*N* = 5)	NR	*R* = 24–40 *M* = NR
	Marvin et al. [41]	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 et al. [85]	Assess reliability and validity of the Murray secretion scale	Patients (*N* = 35)	NR	*R* = NR *M* = NR
Yale pharyngeal residue severity rating scale					
Neubauer et al., 2015 [68]	Neubauer et al. [68]	Develop an image-based scoring system to assess the amount of valleculae and pyriform sinus residue	25 images of FEES from adults attending an urban hospital. 260 ratings completed on these images	NR	*R* = NR *M* = NR
VFSS					
FDS					
Han et al., 2001 [70]	Han et al. [70]	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
VDS					
Han et al., 2008 [71]	Han et al. [71]	Develop a measure to predict long-term prognosis of stroke patients with dysphagia	Patients within 72 h of admission, repeated at 6 months post stroke (*N* = 83)	Stroke	*R* = 38–85 *M* = 62
	Kim et al. [80]	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 et al. [81]	Determine the clinical applicability of the VDS to multiple aetiologies	Patients who underwent VFSS (*N* = 1995)	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

Table 6 continued

Measure; reference	Study on psychometrics	Study purpose	Study population, number (*N*)	Aetiologies, number (*N*)	Age (range, [*R*]) and/or Mean [*M*] years
DIGEST					
Hutcheson et al., 2017 [72]	Hutcheson et al. [72]	Explore feasibility and psychometrics of DIGEST	Patients post treatment ($N = 100$)	Head and neck cancers	$R = 47–84$ $M = 61$
MBSImp					
Martin-Harris et al., 2008 [73]	Martin-Harris et al. [73]	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 et al. [86]	Examine relationship between MBSImp 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%)	$R = 21–94$ $M = 58$
Two variables (duration–bolus transit & volume–residue)					
Daniels et al., 2006 [69]	Daniels et al. [69]	See Daniels, under PAS	RE	RE	RE
Twelve variables					
Frowen et al., 2008 [23]	Frowen et al. [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
Single variable (delay)					
Karnell and Rogus, 2005 [21]	Karnell and Rogus [21]	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 ($N = 1$)	$R = 29.7–83$ $M = 61.6$

Table 6 continued

Measure; reference	Study on psychometrics	Study purpose	Study population, number (*N*)	Aetiologies, number (*N*)	Age (range, [*R*]) and/or Mean [*M*] years
Single variable (residue–location)					
Omari et al., 2011 [74]	Omari et al. [74]	See Omari, under PAS	RE	RE	RE
FEES and VFSS					
UCSF standardised grading form					
Curtis et al., 2016 [75]	Curtis et al. [75]	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
Single variable (residue–volume)					
Park et al., 2015 [76]	Park et al. [76]	See Park, under PAS	RE	RE	RE
PAS					
Rosenbek et al., 1996 [25]	Butler et al. [84]	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 et al. [83]	Determine reliability of the PAS as a function of clinician experience	35 swallow recordings	NR	*R* = NR *M* = NR
	Colodny [78]	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 et al. [69]	Develop a standard method of using VFSS to define dysphagia	Patients (*N* = 9)	Stroke	*R* = 50–78 *M* = 62
			Healthy adults (*N* = 13)	Males with no history of neurological disease, COPD, head and neck cancer or dysphagia	*R* = 54–76 *M* = 64
	Hind et al. [50]	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 et al. [82]	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

Table 6 continued

Measure; reference	Study on psychometrics	Study purpose	Study population, number (*N*)	Aetiologies, number (*N*)	Age (range, [*R*]) and/or Mean [*M*] years
	McCullough et al. [26]	Assess reliability of the PAS	Patients with stroke (*N* = 20)	Stroke within 6 weeks of VFSS	*R* = 40–96 *M* = 67.8
	Omari et al. [74] Park et al. [76]	Determine if bolus residue may be detected without use of VFSS Compare diagnostic efficacy between VFSS and endoscopist-directed FEES	Adults (*N* = 17) and children (*N* = 6) with dysphagia	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
			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 et al. [25]	Define and describe use and development of the PAS and report reliability data	Patients with dysphagia (*N* = 15)	Stroke	*R* = NR *M* = NR

NR not reported, *RE* reported elsewhere

measures also rarely utilised subscales or summed scores. A total of three measures included summed overall scores [FDS [70], VDS [71], Single variable–residue (location), [74]], while two utilised subscales [MBSImp [73] and DIGEST [72]].

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) [25, 64–68] and ranged from 3- to 8-point scales. Two measures used nominal response scales [75, 77]. Conversely, nominal scales were more common among VFSS measures (n = 6) [21, 70–72, 74, 75]. 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 = 3) ranging from 2- to 8-points [25, 69, 73], dichotomous scales (n = 3), and continuous response options such as time (n = 2) were used less frequently in VFSS [21, 69]. Two measures used multiple types of these response options [21, 69].

Table 6 synthesises information from the 27 studies which examined the 18 measures with multiple psychometric data/data other than reliability only. The majority of measures had their psychometrics investigated by only one study (n = 13) [21, 23, 64–66, 68–70, 72, 74–76]. All but one study examined adult populations; one included children and adults [64]. Age varied widely, from 10 to 100 years (mean = 61.4 years; SD = 7.7). Aetiology was similarly varied, 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 were stroke (n = 22 studies) [21, 23, 25, 26, 64–67, 69–71, 73, 74, 76, 78–81], degenerative neurological diseases (n = 14) [21, 41, 50, 66, 69, 74, 76, 77, 79, 81, 82], and head and neck cancers (n = 9) [21, 23, 41, 66, 72, 73, 77, 81, 82]. Number of participants studied ranged from 14 to 1,995 (mean = 161.6 [SD = 376.7]; median = 45 [IQR 80]).

Psychometric Properties

Table 4 summarises the quality ratings of 21 measures where information is available about inter- and or intra-rater reliability only. According to COSMIN ratings, three studies had 'Poor' methodological quality [49, 62, 63], which resulted in the relevant reliability type for that study

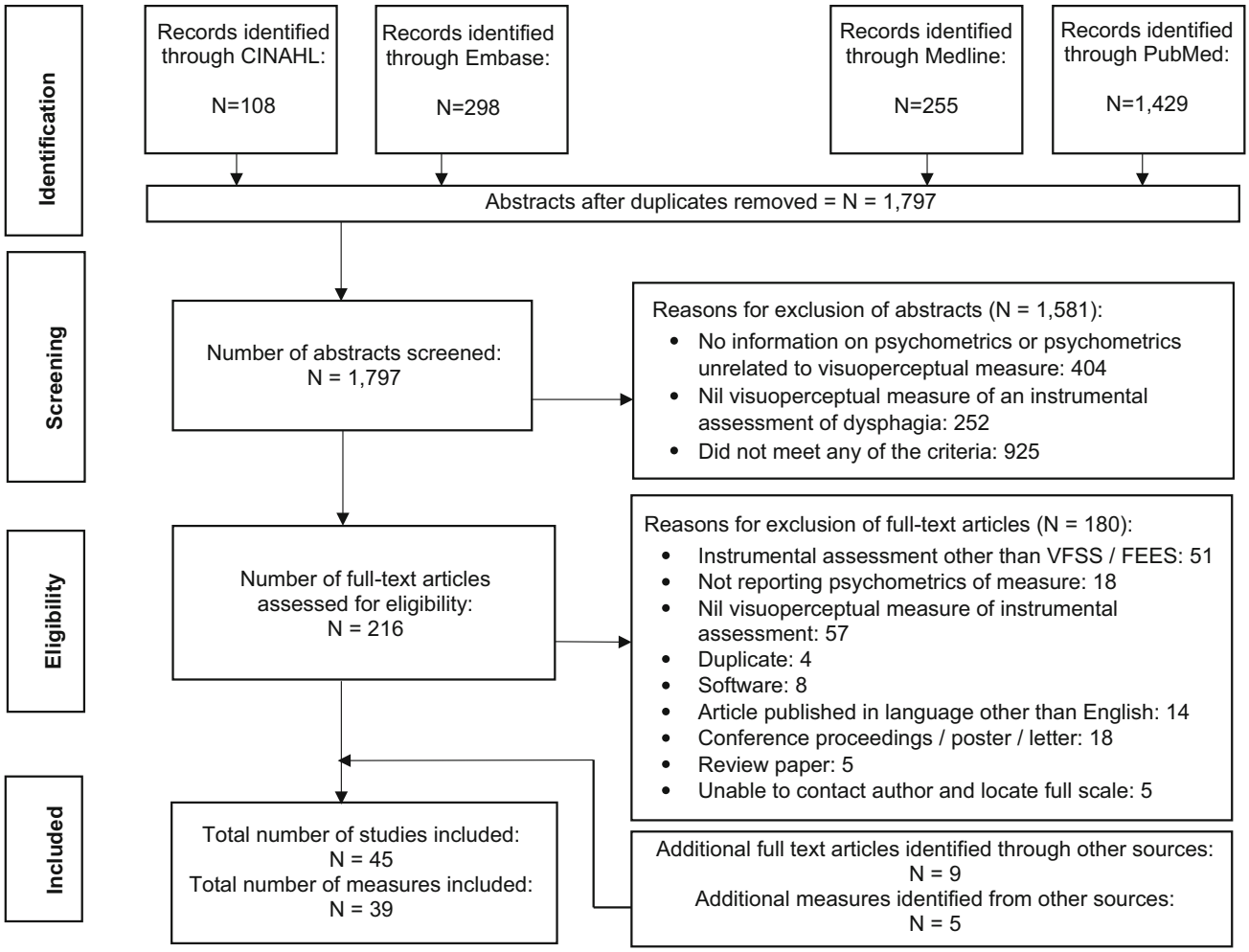

Fig. 2 Flow diagram of reviewing process according to PRISMA [28]

being excluded from further analysis). Eleven studies had 'Fair' [26, 41, 43, 44, 46, 48, 49, 55, 58, 60, 77], thirteen 'Good' [23, 26, 42, 44–47, 50, 51, 53, 56, 61, 62], and one 'Excellent' [42]. The overall quality ratings, based on Terwee et al. [36] and Schellingerhout et al. [38], resulted in two measures with moderate negative ratings [47, 53], two with limited negative [41, 60], two indeterminate [43, 49], five with limited positive [41, 44, 46, 58, 59], four with moderate positive scores [50, 53, 56, 62], and seven with conflicting ratings [42, 45, 48, 51, 54, 55, 61].

Table 7 describes the methodology quality ratings of studies (as determined by COSMIN) which report on more than one psychometric property, or properties other than solely inter-/intra-rater reliability. Among these studies, reliability continues to remain the most common psychometric property reported on (n = 24) [21, 23, 25, 41, 50, 54, 64–66, 68–70, 72, 73, 75, 78–80, 82–85], followed by hypothesis testing (n = 19) [21, 23, 64, 66, 68, 70, 72–76, 81, 85, 86]. In addition, one study reported on internal consistency [66], 12 on content validity [25, 64–69, 71–73, 77], and two on structural validity [66, 73]. 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 [21, 25, 41, 50, 54, 64, 67, 69, 74, 76, 78, 82–85]. Although all studies were published in English, it is likely two measures were developed in another language [65, 76]. 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 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 in study quality (n = 4) [54, 69, 73, 76] were excluded from analysis of psychometric property quality (Table 9).

Table 8 provides a summary of the quality of psychometric properties based on Terwee et al. [36] and Schellingerhout et al. [38] criteria. Finally, Table 9 summarises of the overall quality ratings per psychometric property of nine FEES measures and nine VFSS measures, as evaluated by Schellingerhout et al. [38] criteria; this combines Table 7's COSMIN methodological quality scores with the quality criteria in Table 8. One measure, the PAS [25], assessed both FEES and VFSS; as such, the results were reported separately as it had different psychometric properties for FEES and VFSS. The notes section of Table 9 describes the criteria used to rate the overall psychometric quality. Reliability was the most commonly (n = 14) assessed psychometric property [21, 23, 25, 64–68, 70–72, 75, 77], followed by hypothesis testing (n = 13) [21, 23, 25, 64, 66–68, 70–74] and content validity (n = 12) [25, 64–67, 69, 71–73, 77]. Structural validity was analysed twice [66, 73] and one study reported on internal consistency [66]. Up to five psychometric properties were reported on per measure, with a mode of three per measure. Only eight measures were found to have one or more properties with positive evidence of psychometric soundness [64–66, 68, 70–72, 77]. Four measures had conflicting evidence [21, 25, 72]; these were due to scores on reliability analyses being below the quality cut-off score for ICC or weighted Kappa < 0.7 in one study but meeting the minimum required cut-off score (> 0.7) in another study, where both studies displayed adequate design. One measure had limited negative evidence [71]. 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.

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 images which are produced through these assessments are not meaningful in and of themselves. They 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 [27], which appraises the quality of 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 the examination of the quality of psychometrics of these measures [27, 32]. This systematic review therefore provides a comprehensive summary of the quality of psychometric properties of visuoperceptual measures which are currently available for VFSS and FEES.

Psychometric Quality of Measures Overall

A total of 18 measures reported on more than one psychometric property [21, 23, 25, 64–75, 77], or properties other than reliability only, while 21 measures had studies which reported solely on inter-/intra-rater reliability (Table 4) [23, 26, 41–51, 53, 56, 58, 60, 62, 63, 82]. 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 [66, 73], and one on internal consistency [66] (Tables 7, 8, 9). 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, the Videofluoroscopic Dysphagia scale [71] features 14 components that sum together to create a total score. This measure's structural validity is unknown; this means the measure may not be unidimensional, and thus makes the use of the total score questionable. 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 to other random factors. Inadequate information on this property means it cannot be assumed that alteration in a patient's scores indicate improving or worsening swallow function versus changes other related factors.

The most common overall result across all of the assessed psychometric properties was 'indeterminate' (64%) (Table 9). '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 [21, 23, 25, 64, 66–68, 70–74]. Hypothesis testing examines the relationship of the measure compared to other measures, or difference between groups. COSMIN standards state specific hypotheses should be formulated a priori, with expected direction and magnitude of

Table 7 Overview of the methodological quality assessment results using the COSMIN checklist: studies reporting on psychometric properties of VFSS and FEES measures

Measure and author(s)	Internal consistency[a]	Reliability		Measurement error	Content validity	Structural validity	Hypothesis testing
		Inter	Intra				
FEES							
UCSF standardised grading forms							
Curtis et al. [75]	NR	Fair (27.2%)[d]	Fair (33.3%)[d]	NR	NR	NR	Poor (13.0%)
Marianjoy 3-point secretion severity scale							
Donzelli et al. [64]							
Total scale	N/A	NR	NR	NR	Fair (50.0%)	NR	NR
Penetration	N/A	NR	NR	NR	NR	NR	Fair (30.4%)
Aspiration	N/A	NR	NR	NR	NR	NR	Fair (30.4%)
Diet outcomes	N/A	NR	NR	NR	NR	NR	Fair (29.4%)
Marianjoy 5-point secretion severity scale							
Donzelli et al. [64]							
Total scale	N/A	Fair (27.3%)	NR	NR	Good (71.4%)	NR	NR
Penetration	N/A	NR	NR	NR	NR	NR	Fair (34.8%)
Aspiration	N/A	NR	NR	NR	NR	NR	Fair (34.8%)
Diet outcomes	N/A	NR	NR	NR	NR	NR	Fair (47.1%)
Tracheostomy status	N/A	NR	NR	NR	NR	NR	Fair (47.1%)
Dysphagia score[b]							
Dziewas et al. [65]	NR	Fair (27.2%)	NR	NR	Fair (42.9%)	NR	NR
P score[c]							
Farneti [77]	NR	NR	NR	NR	Good (57.1%)	NR	NR
Farneti et al. [79]	NR	Fair (26.7%)	Fair (42.4%)	NR	NR	NR	NR
BRACS							
Kaneoka et al. [66]	Good (71.4%)	Fair (26.3%)	Good (57.6%)	NR	Good (71.4%)	Good (58.3%)	Fair (39.1%)
Murray secretion scale							
Murray et al. [67]	N/A	NR	NR	NR	Excellent (78.6%)	NR	NR
Marvin et al. [41]	N/A	NR	Fair (31.3%)	NR	NR	NR	NR
Pluschinski et al. [85]	N/A	Fair (26.3%)	Good (54.5%)	NR	NR	NR	Fair (30.4%)
Yale pharyngeal residue severity rating scale							
Neubauer et al. [68]	NR	Good (68.4%)	Excellent (81.8%)	NR	Fair (35.7%)	NR	Fair (43.5%)
Single variable (residue–volume)							
Park et al. [76]	N/A	NR	NR	NR	NR	NR	Poor (21.7%)
PAS–FEES							
Butler et al. [84]	N/A	Excellent (81.8%)	NR	NR	NR	NR	NR
Butler et al. [83]	N/A	Poor (21.1%)	Fair (42.4%)	NR	NR	NR	NR
Colodny [78]	N/A	Fair (31.6%)	Good (54.6%)	NR	NR	NR	NR
Kelly et al. [82]	N/A	Good (52.6%)	Good (60.6%)	NR	NR	NR	NR
Park et al. [76]	N/A	NR	NR	NR	NR	NR	Poor (21.7%)
VFSS							
UCSF standardised grading forms							
Curtis et al. [75]	NR	Poor (10.5%) [d]	Fair (30.3%) [d]	NR	NR	NR	Poor (13.0%)
VDS							
Han et al. [71]	NR	NR	NR	NR	Fair (50.0%)	NR	NR
Kim et al. [81]	NR	NR	NR	NR	NR	NR	Fair (47.8%)
Kim et al. [80]	NR	Fair (31.6%)	NR	NR	NR	NR	NR

Table 7 continued

Measure and author(s)	Internal consistency[a]	Reliability		Measurement error	Content validity	Structural validity	Hypothesis testing
		Inter	Intra				
FDS							
Han et al. [70]	NR	Fair (44.8%)	NR	NR	NR	NR	Fair (30.4%)
DIGEST							
Hutcheson et al. [72]	NR	Good (57.9%)	Good (63.3%)	NR	Excellent (78.6%)	NR	Fair (43.5%)
MBSImp							
Martin-Harris et al. [73]	NR	Poor (10.5%)	Poor (24.2%)	NR	Good (64.3%)	Excellent (83.3%)	Good (65.2%)
Gullang et al. [86]	NR	NR	NR	NR	NR	NR	Fair (30.4%)
Single variable (residue–volume)							
Park et al. [76]	N/A	NR	NR	NR	NR	NR	Poor (21.7%)
PAS–VFSS							
Rosenbek et al. [25]	N/A	Good (52.6%)	Good (66.7%)	NR	Good (57.14%)	NR	NR
Daniels et al. [69]	N/A	Poor (10.5%)	Poor (15.1%)	NR	NR		
Hind et al. [50]	N/A	Fair (36.8%)	NR	NR	NR	NR	
Kelly et al. [82]	N/A	Good (52.6%)	Good (60.6%)	NR	NR	NR	
McCullough et al. [54]	N/A	Poor (21.1%)	Fair (42.4%)	NR	NR	NR	
Omari et al. [74]	N/A	NR	NR	NR	NR	NR	
Park et al. [76]	N/A	NR	NR	NR	NR	NR	Poor (21.7%)
Two variables (duration and volume–residue)							
Daniels et al. [69]							
Bolus duration (s)	NR	Poor (15.4%)	Poor (24.1%)	NR	Fair (35.7%)	NR	NR
Bolus clearance	NR	Poor (10.5%)	Poor (15.1%)	NR	Fair (28.6%)	NR	NR
Twelve variables							
Frowen et al. [23]							
Semi-solids	NR	Good (79.0%)	Good (57.6%)	NR	NR	NR	Good (60.1%)
Liquids	NR	Fair (26.3%)	Good (57.6%)	NR	NR	NR	Good (60.1%)
Single variable (delay)							
Karnell and Rogus [21]							
Latency (s)	N/A	Good (66.7%)	Good (64.0%)	NR	NR	NR	Fair (39.1%)
Dichotomous options	N/A	Good (57.9%)	Good (63.6%)	NR	NR	NR	Fair (39.1%)
Severity	N/A	Good (57.9%)	Good (63.6%)	NR	NR	NR	NR
Single variable (residue)							
Omari et al. [74]	NR	NR	NR	NR	NR	NR	Fair (47.8%)

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 8 Quality of psychometric properties per study based on the criteria by Terwee et al. [36] and Schellingerhout [38]

Measure and author(s)	Internal consistency	Reliability		Measurement error	Content validity	Structural validity	Hypothesis testing
		Intra	Inter				
FEES							
UCSF standardised grading forms							
Curtis et al. [75]	NR	?	?	NR	NR	NR	NE
Marianjoy 3-point secretion severity scale							
Donzelli et al. [64]	N/A	NR	NR	NR	?	NR	?
Marianjoy 5-point secretion severity scale							
Donzelli et al. [64]	N/A	+	NR	NR	?	NR	?
Dysphagia score							
Dziewas et al. [65]	NR	+	NR	NR	+	NR	NR
P score							
Farneti [77]	NR	NR	NR	NR	?	NR	NR
Farneti et al. [79]	NR	+	+	NR	NR	NR	NR
BRACS							
Kaneoka et al. [66]	?	+	+	NR	?	+	?
Yale pharyngeal residue severity rating scale							
Neubauer et al. [68]	NR	+	+	NR	?	NR	?
Murray secretion scale							
Murray et al. [67]	N/A	NR	NR	NR	?	NR	NR
Marvin et al. [41]	N/A	NR	?	NR	NR	NR	NR
Pluschinski et al. [85]	N/A	?	?	NR	NR	NR	?
Single variable (residue–volume)							
Park et al. [76]	N/A	NR	NR	NR	NR	NR	NE
PAS–FEES							
Butler et al. [84]	N/A	−	NR	NR	NR	NR	NR
Butler et al. [83]	N/A	+	+	NR	NR	NR	NR
Colodny [78]	N/A	±	+	NR	NR	NR	NR
Kelly et al. [82]	N/A	−	+	NR	NR	NR	NR
Park et al. [76]	N/A	NR	NR	NR	NR	NR	NE
VFSS							
UCSF standardised grading forms							
Curtis et al. [75]	NR	NE	?	NR	NR	NR	NE
VDS							
Han et al. [71]	NR	NR	NR	NR	+	NR	NR
Kim et al. [81]	NR	NR	NR	NR	NR	NR	?
Kim et al. [80]	NR	−	NR	NR	NR	NR	NR
FDS							
Han et al. [70]	NR	+	NR	NR	NR	NR	?
DIGEST							
Hutcheson et al. [72]	NR	−	+	NR	+	NR	?
MBSImp							
Martin-Harris et al. [73]	NR	NE	NR	NR	?	?	?
Gullang et al. [86]	NR	NE	NR	NR	NR	NR	?
Single variable (residue–volume)							
Park et al. [76]	N/A	NR	NR	NR	NR	NR	NE
PAS–VFSS							
Daniels et al. [69]	N/A	NE	NE	NR	NR	NR	NR
Hind et al. [50]	N/A	+	NR	NR	NR	NR	NR

Table 8 continued

Measure and author(s)	Internal consistency	Reliability		Measurement error	Content validity	Structural validity	Hypothesis testing
		Intra	Inter				
Kelly et al. [82]	N/A	−	+	NR	NR	NR	NR
McCollough et al. [26]	N/A	±	?	NR	NR	NR	NR
Omari et al. [74]	N/A	NR	NR	NR	NR	NR	?
Park et al. [76]	N/A	NR	NR	NR	NR	NR	NE
Rosenbek et al. [25]	N/A	±	±	NR	?	NR	NR
Two variables (duration–bolus transit and volume–residue)							
Daniels et al. [69]							
Bolus Duration	NR	NE	NE	NR	?	NR	NR
Residue	NR	NE	NE	NR	?	NR	NR
Twelve variables							
Frowen et al. [23]							
Semi-Solids	NR	?	?	NR	NR	NR	?
Liquids	NR	?	?	NR	NR	NR	?
Single variable (delay)							
Karnell and Rogus [21]							
Latency (s)	N/A	?	?	NR	NR	NR	?
Dichotomous options	N/A	±	±	NR	NR	NR	?
Severity	N/A	−	−	NR	NR	NR	NR
Single variable (residue—location)							
Omari et al. [74]	NR	NR	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)

correlations stated [32]. 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. It should also be noted that according to the COSMIN taxonomy, recruitment of more than 100 participants is 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 (Table 6). Where validation studies use a limited sample size, the accuracy of their conclusions and the generalisability of results to the wider population is questionable.

Content validity was another psychometric property with high rates of 'indeterminate' findings (Table 9). 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 [32]. The content validity ratings of measures included in this review were negatively affected by one or all of the following short-comings: 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. For example, the development of Secretion Severity Scale [67] involved an insufficient number and range of experts. The literature was reviewed, but this process was not described, nor was the interview with experts. Description of how concepts were operationalised was also lacking. Deficiencies in establishing and reporting on content validity have 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'

Table 9 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						
UCSF Standardised Grading Form Curtis et al. (75)	NR	Indeterminate	NR	NR	NR	NE
Marianjoy 3-Point secretion severity scale Donzelli et al. (64)	N/A	NR	NR	Indeterminate	NR	Indeterminate
Marianjoy 5-Point secretion severity scale Donzelli et al. (64)	N/A	Limited (positive)	NR	Indeterminate	NR	Indeterminate
Dysphagia Score Dziewas et al. (65)	NR	Limited (positive)	NR	Limited (positive)	NR	NR
P-Score Farneti (77)	NR	Limited (positive)	NR	Indeterminate	NR	NR
BRACS Kaneoka et al. (66)	Indeterminate	Limited (positive)	NR	Indeterminate	Moderate (positive)	Indeterminate
Murray Secretion Scale Murray et al. (67)	N/A	Indeterminate	NR	Indeterminate	NR	Indeterminate
Yale Pharyngeal Residue Severity Rating Scale Neubauer et al. (68)	NR	Strong (positive)	NR	Indeterminate	NR	Indeterminate
PAS Rosenbek et al. (25)	N/A	Conflicting	NR	NR	NR	NE
VFSS						
VDS Han et al. (71)	NR	Limited (negative)	NR	Limited (positive)	NR	Indeterminate
FDS Han et al. (70)	NR	Limited (positive)	NR	NR	NR	Indeterminate
DIGEST Hutcheson et al. (72)	NR	Conflicting	NR	Strong (positive)	NR	Indeterminate
MBSImp Martin-Harris et al. (73)	NR	NE	NR	Indeterminate	Indeterminate	Indeterminate
PAS Rosenbek et al. (25)	N/A	Conflicting	NR	Indeterminate	NR	Indeterminate
Two Variables (Duration and Volume) Daniels et al. (69)	NR	NE	NR	Indeterminate	NR	NR
Twelve variables Frowen et al. (23)	NR	Indeterminate	NR	NR	NR	Indeterminate
Single variable (Delay) Karnell & Rogus (21)	N/A	Conflicting	NR	NR	NR	Indeterminate
Single variable (Residue - location) Omari (74)	NR	NR	NR	NR	NR	Indeterminate

Single variable—Residue (volume) Park et al. [76] and VFSS reliability of UCSF standardised grading form excluded from final analysis (Table 9) due to 'Poor' scores (Table 7)

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

results, 'limited' strength of evidence was also a frequent finding (17%) [64–66, 70, 71, 77]. This was the result of the low number of psychometric properties investigated per study for each measure; most measures (31 of the 39 measures) had only one study which investigating a very limited range of psychometric properties. 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 measure most commonly included variables related to residue, penetration/aspiration, and secretions (Table 5). This is likely a reflection of seminal works on the use and analysis of the FEES and VFSS [52, 87], and the importance of aspiration as a predictor of aspiration pneumonia and chronic dysphagia [88, 89].

None of the studies described how response options were designed or decisions on the number of items were made. Measure design may have had an impact on the quality of psychometric properties; the analysis of overall scores of measures with solely reliability data revealed that use of fewer items appeared to correspond with positive overall scores. It was also noted VFSS measures on average used three more items than FEES measures, and the upper range of items used was higher (23 for VFSS vs. 16 for FEES). 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 (Tables 5, 6, 7, 8, 9), only seven utilised subscales and/or summed scores [66, 70–74, 77]. 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 four utilised paediatric populations [49, 64, 74, 77] (Table 6). 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 [90]. 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) that the person statistic (i.e. observed score) is item sample dependent; and (b) the item statistics are examinee/person sample dependent. This poses some difficulties in CTT's application in some measurement situations [91]. 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 [92]. 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 [93, 94]. 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 the 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 (Table 9). 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 Boston Residue and Clearance Scale (BRACS) [66] and the Dysphagia score [65]; BRACS had limited positive evidence for reliability and moderate positive for structural validity, while the Dysphagia score had limited positive evidence of reliability and content validity. As information about only two measurement properties was 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. The psychometric properties of internal consistency, measurement error, structural validity, and hypothesis testing should be investigated in future research.

In terms of VFSS analysis, the Dynamic Imaging Grade of Swallow Toxicity (DIGEST) [72] had the highest rated evidence overall, with strong positive evidence for content validity. An indeterminate score was recorded in hypothesis testing due to lack of a priori hypotheses, and conflicting reliability was found due to strong intra-rater reliability but weak inter-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 in any psychometric property. Of the measures with reliability data only, an unnamed 'presence/absence of aspiration' dichotomous scale [50], an unnamed scale of temporal and spatial variables [62], and an unnamed scale of temporal variables [56] 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 FEES 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. Articles included in this review are limited to studies that were found based on the search strategies as described in Supplementary Table 1, and hand-searching references of accepted articles. 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 that 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 researchers 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 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 were 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 have 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 hence suitability for use in practice and research settings, may be limited or 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 adopt appropriate study designs and reporting practices.

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

Author Contributions The authors KS, RC, TB and RS have conceived, designed, and performed the experiment, contributed materials/analysis tools and to the writing of the paper. The authors KS and RS have analysed the data.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Dodrill P, Gosa MM. Pediatric dysphagia: physiology, assessment, and management. Ann Nutr Metab. 2015;66(Suppl. 5):24–31.
2. Kalf J, De Swart B, Bloem B, Munneke M. Prevalence of oropharyngeal dysphagia in Parkinson's disease: a meta-analysis. Parkinsonism Relat Disord. 2012;18(4):311–5.
3. 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. Support Care Cancer. 2015;23(11):3249–55.
4. 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.
5. 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.
6. Bhattacharyya N. The prevalence of dysphagia among adults in the United States. Otolaryngol-Head Neck Surg. 2014;151(5):765–9.
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. Dis Esophagus. 2011;24(7):476–80.
8. Park Y-H, Han H-R, Oh B-M, Lee J, Park J-a YuSJ, et al. Prevalence and associated factors of dysphagia in nursing home residents. Geriatr Nurs. 2013;34(3):212–7.
9. Cichero JA, Heaton S, Bassett L. Triaging dysphagia: nurse screening for dysphagia in an acute hospital. J Clin Nurs. 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. Am J Phys Med Rehabil. 2013;92(6):496–503.
11. DiBardino DM, Wunderink RG. Aspiration pneumonia: a review of modern trends. J Crit 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. Clin Nutr. 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. J Psychosom Res. 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. Int J Stroke. 2010;5(6):466–76.
17. Carnaby-Mann G, Lenius K. The bedside examination in dysphagia. Phys Med Rehabil Clin N Am. 2008;19(4):747–68.
18. McCullough G, Rosenbek J, Wertz R, McCoy S, Mann G, McCullough K. Utility of clinical swallowing examination measures for detecting aspiration post-stroke. J Speech Lang Hear Res. 2005;48(6):1280–93.
19. Langmore SE. History of fiberoptic endoscopic evaluation of swallowing for evaluation and management of pharyngeal dysphagia: changes over the years. Dysphagia. 2017;32(1):27–38.
20. Huckabee M-L, Macrae P, Lamvik K. Expanding instrumental options for dysphagia diagnosis and research: ultrasound and manometry. Folia Phoniatr Logop. 2015;67(6):269–84.
21. Karnell MP, Rogus NM. Comparison of clinician judgments and measurements of swallow response time: a preliminary report. J Speech Lang Hear Res. 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 Med Educ. 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. Nat Rev Gastroenterol Hepatol. 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. Qual Life Res. 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.
29. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–9.
30. 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.
31. Newman RD, Nightingale J. Improving patient access to videofluoroscopy services: role of the practitioner-led clinic. Radiography. 2011;17(4):280–3.
32. 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 Med Res Methodol. 2010;10(1):22.
33. Higgins JP, Green S. Cochrane Handbook for systematic reviews for interventions. New York: Wiley; 2008.
34. Centre for Reviews Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. Layerthorpe, York: CRD University of York; 2009.
35. 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. J Clin Epidemol. 2010;63:737–45.
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 questionnaires. J Clin Epidemiol. 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):e0132299.
38. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet H, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res. 2012;21:659–70.
39. Cordier R, Chen Y, Speyer R, Totino R, Doma K, Leicht A, et al. Child-report measures of occupational performance: a systematic review. PLoS ONE. 2016;11(1):1–24.
40. Doma K, Speyer R, Leicht A, Cordier R. Comparison of psychometric properties between usual-week and past-week self-reported physical activity questionnaires: a systematic review. Int J Behav Nutr Act. 2016. https://doi.org/10.1186/s12966-017-0470-6.
41. Marvin S, Gustafson S, Thibeault S. Detecting aspiration and penetration using FEES with and without food dye. Dysphagia. 2016;31(4):498–504.
42. 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.
43. Rodriguez KH, Roth CR, Rees CJ, Belafsky PC. Reliability of the pharyngeal squeeze maneuver. Ann Otol Rhinol Laryngol. 2007;116(6):399–401.
44. 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. Am J Phys Med Rehabil. 2015;94(1):38–43.
45. 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. J Oral Rehabil. 2010;37(12):884–91.
46. 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 Relat Disord. 2016;28:100–6.
47. Bryant KN, Finnegan E, Berbaum K. VFS interjudge reliability using a free and directed search. Dysphagia. 2012;27(1):53–63.
48. Gibson E, Phyland D, Marschner I. Rater reliability of the modified barium swallow. Aust J Hum Commun Disord. 1995;23(2):54–60.
49. 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.
50. 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.
51. Lee JW, Randall DR, Evangelista LM, Kuhn MA, Belafsky PC. Subjective assessment of videofluoroscopic swallow studies. Otolaryngol-Head Neck Surg. 2017;156(5):901–5.
52. Leonard R, Kendall K. Dysphagia assessment and treatment planning: a team approach. Boston: Cengage Learning; 1997.
53. Mann G, Hankey GJ, Cameron D. Swallowing disorders following acute stroke: prevalence and diagnostic accuracy. Cerebrovasc Dis. 2000;10(5):380–6.
54. 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. Am J Speech-Lang Pathol. 1999;8(2):149–63.
55. Miles A. Inter-rater reliability for speech–language therapists' judgement of oesophageal abnormality during oesophageal visualization. Int J Lang Commun Disord. 2016. https://doi.org/10.1111/1460-6984.12283.
56. 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.
57. Price GJ, Jones CJ, Charlton RA, Allen C. A combined approach to the assessment of neurological dysphagia. Clin Otolaryngol. 1987;12(3):197–201.
58. 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. Int J Otolaryngol. 2015. https://doi.org/10.1155/2015/780197.
59. Scott AG. The development of a scale to assess swallowing function in motor neuron disease using videofluoroscopic techniques. Melbourne: La Trobe University; 1999.
60. Stoeckli SJ, Huisman TA, Seifert BA, Martin-Harris BJ. Interrater reliability of videofluoroscopic swallow evaluation. Dysphagia. 2003;18(1):53–7.
61. 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? Clin Otolaryngol. 2006;31(5):425–32.
62. Nordin NA, Miles A, Allen J. Measuring competency development in objective evaluation of videofluoroscopic swallowing studies. Dysphagia. 2017;32(3):427–36.
63. 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.
64. 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 Otol Rhinol Laryngol. 2003;112(5):469–75.
65. 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. Cerebrovasc Dis. 2008;26(1):41–7.
66. 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 Phoniatr Logop. 2013;65(6):312–7.
67. 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.
68. 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.
69. Daniels SK, Schroeder MF, McClain M, Corey DM. Dysphagia in stroke: development of a standard method to examine swallowing recovery. J Rehabil Res Dev. 2006;43(3):347.
70. Han TR, Paik N-J, Park JW. Quantifying swallowing function after stroke: a functional dysphagia scale based on videofluoroscopic studies. Arch Phys Med Rehabil. 2001;82(5):677–82.
71. 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.
72. 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.
73. 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.
74. 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. Am J Gastroenterol. 2011;106(10):1796–802.
75. 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.
76. 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 Liver. 2015;9(5):623.
77. Farneti D. Pooling score: an endoscopic model for evaluating severity of dysphagia. Acta Otorhinolaryngol Ital. 2008;28(3):135.
78. 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.
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 Otorhinolaryngol Ital. 2014;34(2):105.
80. Kim DH, Choi KH, Kim HM, Koo JH, Kim BR, Kim TW, et al. Inter-rater reliability of videofluoroscopic dysphagia scale. Ann Rehab Med. 2012;36(6):791–6.
81. 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.
82. 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.
83. Butler SG, Markley L, Sanders B, Stuart A. Reliability of the penetration aspiration scale with flexible endoscopic evaluation of swallowing. Ann Otol Rhinol Laryngol. 2015;124(6):480–3.
84. 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. Ann Otol Rhinol Laryngol. 2011;120(5):288–95.
85. Pluschinski P, Zaretsky E, Stöver T, Murray J, Sader R, Hey C. Validation of the secretion severity rating scale. Eur Arch Otorhinolaryngol. 2016;273(10):3215–8.
86. 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. Ann Otol Rhinol Laryngol. 2012;121(11):738–45.
87. Logemann JA. Manual for the videofluorographic study of swallowing. Austin: Pro ed; 1993.
88. 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. J Stroke Cerebrovas Dis. 2012;21(7):569–76.
89. van der Maarel-Wierink CD, Vanobbergen JN, Bronkhorst EM, Schols JM, de Baat C. Meta-analysis of dysphagia and aspiration pneumonia in frail elders. J Dent Res. 2011;90(12):1398–404.
90. Streiner DL, Norman GR, Cairney J. Item response theory. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 2014.
91. Fan X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educ Psychol Measur. 1998;58(3):357.
92. Duong M. Introduction to item response theory and its applications. Michigan: Michigan State University; 2004.
93. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16:5–18.
94. Reise SP, Ainsworth AT, Haviland MG. Item response theory: fundamentals, applications, and promise in psychological research. Curr Dir Psychol Sci. 2005;14(2):95–101.
95. 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. J Clin Epidemiol. 2010;63(7):737–45.
96. Leonard RJ, Kendall KA, McKenzie S, Gonçalves MI, Walker A. Structural displacements in normal swallowing: a videofluoroscopic study. Dysphagia. 2000;15(3):146–52.

Katina Swan BSpPath(Hons)

Reinie Cordier PhD

Ted Brown PhD

Renée Speyer PhD

Affiliations

Katina Swan[1] · Reinie Cordier[1] · Ted Brown[2] · Renée Speyer[1,3,4]

[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