Dysphagia (2017) 32:250–260
DOI 10.1007/s00455-016-9754-2

ORIGINAL ARTICLE

Evaluating the Psychometric Properties of the Eating Assessment Tool (EAT-10) Using Rasch Analysis

R. Cordier[1,2] · A. Joosten[1] · P. Clavé[3,4] · A. Schindler[5] · M. Bülow[6,7,8] · N. Demir[9] · S. Serel Arslan[9] · R. Speyer[2,10]

Received: 3 May 2016 / Accepted: 18 October 2016 / Published online: 21 November 2016

Abstract Early and reliable screening for oropharyngeal dysphagia (OD) symptoms in at-risk populations is important and a crucial first stage in effective OD management. The Eating Assessment Tool (EAT-10) is a commonly utilized screening and outcome measure. To date, studies using classic test theory methodologies report good psychometric properties, but the EAT-10 has not been evaluated using item response theory (e.g., Rasch analysis). The aim of this multisite study was to evaluate the internal consistency and structural validity and conduct a preliminary investigation of the cross-cultural validity of the EAT-10; floor and ceiling effects were also checked. Participants involved 636 patients deemed at risk of OD, from outpatient clinics in Spain, Turkey, Sweden, and Italy. The EAT-10 and videofluoroscopic and/or fiberoptic endoscopic evaluation of swallowing were used to confirm OD diagnosis. Patients with esophageal dysphagia were excluded to ensure a homogenous sample. Rasch analysis was used to investigate person and item fit statistics, response scale, dimensionality of the scale, differential item functioning (DIF), and floor and ceiling effect. The results indicate that the EAT-10 has significant weaknesses in structural validity and internal consistency. There are both item redundancy and lack of easy and difficult items. The thresholds of the rating scale categories were disordered and gender, confirmed OD, and language, and comorbid diagnosis showed DIF on a number of items. DIF analysis of language showed preliminary evidence of problems with cross-cultural validation, and the measure showed a clear floor effect. The authors recommend redevelopment of the EAT-10 using Rasch analysis.

Keywords Measurement · Classic test theory · Item response theory · Reliability · Validity · Oropharyngeal dysphagia · Deglutition · Deglutition disorders

Electronic supplementary material The online version of this article (doi:10.1007/s00455-016-9754-2) contains supplementary material, which is available to authorized users.

✉ R. Cordier
reinie.cordier@curtin.edu.au

[1] School of Occupational Therapy and Social Work, Curtin University, Perth, WA, Australia

[2] College of Healthcare Sciences, James Cook University, Townsville, QLD, Australia

[3] Unitat d'Exploracions Funcionals Digestives, Department of Surgery, Hospital de Mataró, Universitat Autònoma de Barcelona, Mataró, Spain

[4] Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto de Salud Carlos III, Barcelona, Spain

[5] Phoniatric Unit, Department of Biomedical and Clinical Sciences "L. Sacco", Università degli Studi di Milano, Milan, Italy

[6] Diagnostic Centre of Imaging and Functional Medicine, Malmö, Sweden

[7] Department of Clinical Sciences, Lund University, Malmö, Sweden

[8] Skane University Hospital Malmö, Malmö, Sweden

[9] Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Hacettepe University, Ankara, Turkey

[10] Department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Center, Leiden, The Netherlands

Introduction

Oropharyngeal dysphagia (OD) is an underdiagnosed swallowing disorder that can cause severe nutritional and respiratory problems, and it impacts other domains of patient health [1]. Health consequences can include increased risk of dehydration, malnutrition, aspiration pneumonia, and death [2, 3]. OD can also impact upon domains of patient's health-related quality of life and well-being [4]. For example, dysphagia can limit social opportunities and mealtime pleasures [4], and be associated with anxiety, distress, and isolation during mealtimes [5]. OD is a highly prevalent disorder, affecting both the general and clinical populations [1].

Prevalence data are widely variable within different clinical groups as well as within the general population. In the general population, OD prevalence varies between 2.3 and 16 % [6]. Prevalence estimates for selected diagnostic clinical populations range widely: 8.1–80 % in acute stroke; 11–60 % in Parkinson's disease; and approximately 30 % in people with traumatic brain injury [7]. This variation in the prevalence of OD is due to a range of factors, including discrepancies in the characteristics of the study populations or stage of underlying disease progression between studies, the lack of a universally accepted definition of dysphagia, and inconsistent use of screening or assessment tools and outcome measures [7].

Inconsistency in screening and assessment is concerning, as early and reliable screening for OD symptoms in at-risk populations is a crucial first stage in effective OD management [8]. Patients that fail screening require further assessment. Currently, videofluoroscopic (VFS) and/or fiberoptic endoscopic evaluation of swallowing (FEES) of swallowing are mooted as the 'gold-standard' assessment of OD [9]. Unfortunately, it is not feasible to perform these gold standard procedures on all patients at risk for OD, because they require specific equipment which is not guaranteed to be available in every healthcare facility [1].

When selecting an appropriate screening measure, it is important to consider whether the purpose of the measure fits the context within which it will be used. Screening is designed to be an initial procedure, utilized to determine who is eligible for further assessment based on the presence of markers for a particular condition, disease, or illness [10]. Depending on the practice context, the degree of discrimination required at screening level will differ. It is therefore important that tools are evaluated in regard to the psychometric properties most essential to their function as a screening tool, to enable decision-making for diagnostic use [11].

Evaluation of screening tools should include reference to sensitivity, specificity, responsiveness, and positive and negative predictive values. Sensitivity and specificity data enable consideration of the overall identification accuracy of an assessment, the ability of an assessment to accurately diagnose the presence or absence of a condition [11]. Identification accuracy indicates the preciseness of the tool in making a diagnosis and can therefore be argued to be the most critical psychometric property when selecting screening tools [11]. To truly be a quality screening measure, the screening tool must also have a high rate of detection for true positives, and low detection rates for false positives. Evaluation of positive predictive values determines level of tool accuracy compared with that of the gold standard in the field [12]: within the context of screening for OD, the success rate of the selected screening tool in identifying those who have OD from those who do not, compared to the use of VFS and FEES.

To effectively screen and diagnose, OD necessitates health professionals to undergo appropriate training to ensure the safe and reliable administration of screening procedures and interpretation of screening results. This is problematic, particularly in rural and remote health services, or in less acute settings such as aged-care facilities. These services therefore require measures relating to swallowing that are both valid and reliable and enable effective clinical management of OD while requiring minimal resources and training [13]. Consequently, patient self-evaluation questionnaires could be a viable alternative for screening patients at risk of OD in practice areas where resources may be limited. Self-evaluation questionnaires typically consist of functional health status (FHS) and health-related quality of life (HR-QoL) components [13]. Key differences in the emphasis of FHS and HR-QoL instruments exist. FHS is concerned with the influence of a given disease on particular functional aspects of a patient's health, such as the ability to perform tasks in multiple domains [14, 15]. Quality of life (QoL), defined by World Health Organization as "…a state of complete physical, mental and social well-being, not merely the absence of disease or infirmity," is increasingly recognized and used as an outcome measure of effects of medical conditions [16]. In conceptual models and related instruments, QoL is commonly operationalized as HR-QoL. HR-QoL refers to aspects of quality of life that impact an individual's health, both physical and mental [17], and is broader in its scope than FHS.

The process of evaluating such measurement scales is commonly underpinned by the application of classic test theory (CTT) or item response theory (IRT) [12]. Both theories have inherent advantages and disadvantages. CTT is relatively simple in terms of procedures and interpretation; however, judgements can only able made on the performance of the test as a whole and the specific sample group it is trialed with. In comparison, IRT assesses the reliability of each item in terms of its contribution to the overall construct being measured and is independent from the testing group. Use of the Rasch measurement model, a type of IRT model, enables further evaluation of the unidimensionality of the scale in determining if item responses are indeed measuring a single trait [18].

While the Eating Assessment Tool (EAT-10) was originally developed to be an outcome measure, the original authors [19] and other authors [1, 9] subsequently suggested a cut-off score for its use as a screening measure in clinical settings [19]. It is a self-report measure assessing a patients' own evaluation of being at risk for dysphagia by looking predominantly at FHS, with a few items related to HR-QoL. Items are scored on a 5-point scale ($0 =$ no problem to $4 =$ severe problem), and item scores are summed to give a possible total score ranging from 0 to 40. Belafsky et al. [19] first suggested a score of three or more to be suggestive of a patient being at risk of swallowing problems and in need of further evaluation.

The literature describes the EAT-10 as a valid, reliable tool with high internal consistency (Cronbach's $\alpha = 0.96$) and intra-item correlations ranging from 0.72 to 0.91 [1, 19–21]. However, when the psychometric properties of the EAT-10 were rated against the international standards for psychometric quality health status measurements [22], all reported properties were rated as "poor," with the authors citing insufficient statistical testing, methodology, and reporting as cause for concern [13]. Furthermore, concerns have been raised about the optimal cut-off score for identifying patients at risk of OD. In evaluating the identification accuracy of the EAT-10, Rofes et al. [1] recommend adjusting the normative cut-off score to 2 to increase sensitivity (0.89), without impacting on specificity (0.82).

The use of a particular tool to evaluate a patient's current health status when screening for OD can only be justified if it has demonstrated reliability and validity. Prior to this study, the psychometric properties of the EAT-10 have consistently been evaluated using CTT, which is problematic for a number of reasons. The application of CTT sees the test as the unit of analysis and as such it is assumed that all items are evaluating the same underlying characteristic [23]. To date, no robust analysis has been conducted on the items of EAT-10 to determine whether the items are reflective of the construct measured, and as such the saliency of its items is unknown. Furthermore, being a screener, it is desirable for the EAT-10 to be short and quick to administer. However, the reliability of results analyzed using CTT is decreased with a decrease in the number of test items [23]. Evaluating the psychometric properties of the EAT-10 through the use of IRT could address these concerns. In IRT, the item is the unit of analysis, which allows for an item-by-item focus for shorter scales [23]. Through IRT, the fit of each item to the construct being measured can be determined using more sophisticated means, and the reliability of results is not impacted upon by a small numbers of items.

This study aimed to evaluate the reliability, validity, and identification accuracy of the EAT-10 using Rasch analysis. In particular, we aimed to evaluate the item and person fit characteristics, the response scale, the dimensionality of the scale, differential item function, and if the EAT-10 displayed floor and ceiling effects.

Methods

Participants

Six academic hospitals provided retrospective data on patients at risk for OD. All data were collected consecutively during patients visiting outpatient clinics of dysphagia or otorhinolaryngology at the Hacettepe University (Turkey), Hospital de Mataró (Spain), Sacco Hospital in Milan (Italy), and Skane University Hospital Malmö (Sweden). Patients with severe cognitive problems were excluded. To ensure maximum homogeneity of the sample, patients with esophageal dysphagia were excluded. Only those in the clinical population deemed at risk of OD, and who had received gold standard assessment, consisting of VFS and/or FEES and EAT-10 screening, were included.

Protocol

All patients completed the EAT-10 after which a VFS or FEES recording of swallowing was performed as part of standard clinical practice or usual care. The diagnosis of OD was confirmed or repudiated by an experienced speech and language pathologist and/or laryngologist based on VFS and/or FEES, gold standards in the diagnosis of OD [9], that supported the presence of signs such as aspiration, penetration, and residue. Patient characteristics were collected on both gender and age.

The original version of the EAT-10 by Belafsky et al. [19] was published in English (see Supplementary File of the full scale with complete item descriptors). This study used translations of the EAT-10 into four different languages: Turkish, Spanish, Italian, and Swedish. The translated versions were the result of multiple forward and backward translations. English native speakers were involved in the process, as well as native speakers for all languages. Final translations were checked by a team of clinical experts in the field of dysphagia and trialed by pretesting in patients at risk for OD to check the ease of comprehension, the interpretation, and cultural relevance of the EAT-10 items.

Statistical Analysis

Person and Item Fit Statistics

Data were analyzed using Winsteps version 3.92.0 [24]. Fit statistics were used to identify mis-fitting items and the pattern of responses for each person to examine whether the scale was a valid measure of the construct [25]. In this study, interpretation of fit statistics, reported as logits (log odd units) indicate whether the items contributed to a diagnosis of OD and the extent to which the responses of any one person are reliable [25]. Logits reflect item difficulty with lower items in this measure indicating they were more likely to contribute to the diagnosis. Ideal fit is indicated by a MnSq value of 1.0 with an *infit* and *outfit* range of 0.7–1.4 and Z-Standard (Z-STD) score of <2 [18]. Item fit outside the range indicates that the items do not contribute to the construct, and person fit outside the range indicates ratings that are too predictable or too erratic [25]. The item reliability index is a measure (0–1) of internal consistency, and the person reliability index is a measure of replicability of person placement if given different items measuring the same construct [25], with values greater than 0.81 indicating good reliability. The person separation index indicates whether the measure can separate the people into a number of significantly different levels of the trait being measured, and at least two levels are desirable.

Response Scale

Responses across rating categories (i.e., 0–4 rating scale options) were examined for uniform distribution to determine the extent to which the respondents correctly used the response scale. Rasch analysis converts the total EAT-10 score to an average measure score for the item category. Average measure scores (frequency use) are used to reflect whether the rating scale worked effectively, that is, as the category goes up so should the average measure score [26]. Category ordering is indicated by monotonic advances in the average measures. The extent of category disordering is indicated by fit mean squares 0.7–1.4 and Z-STD score of <2 indicating that category is providing misinformation, and consideration should be given to collapsing it with an adjacent category [25].

Step calibrations or Rasch–Andrich thresholds (where there is a 50 % chance of an individual being scored in either category) should progress monotonically to indicate that there is no overlap in categories and they reflect the distance between the categories. The average measure increase should be at least 1.0 logit (on a 5-category scale) to indicate distinct categories, but an increase of >5.0 logits would indicate gaps in the variable [26]. Non-uniformity in responses can also be due to the inclusion of items that do not measure the construct, or poorly defined scale categories.

Dimensionality of the Scale

Following examination to identify potentially mis-fitting persons and items, principal components analysis (PCA) of residuals using Winsteps 3.92.0 [24] was conducted. In contrast to traditional factor analysis, Winsteps conducts a PCA of residuals, not of the original observations [25]. Residuals are the difference between the observed and the expected measure scores, and rather than show loadings on one factor which it shows contrasts between opposing factors [27]. A PCA of residuals looks for patterns in the unexpected data to see if items group together. In PCA of residuals, we are trying to falsify the hypothesis that the residuals are random noise by finding the component that explains the largest possible amount of variance in the residuals, expressed as the first contrast (i.e., first PCA component in the correlation matrix of the residuals) [18]. The Rasch model requires that the scale demonstrates a single construct or unidimensionality on a hierarchical continuum [25]. If the first contrast eigenvalue is small (less than 2 item strength), it is usually regarded as noise, and an eigenvalue of 3 (3 item strength) identifies systematic variance indicative of a second dimension [18]. The person–item dimensionality map provides a schematic representation of the alignment between person ability and item difficulty.

Differential Item Functioning Analysis

Differential item functioning (DIF) analysis is conducted to examine whether the scale items are used the same way by all groups. When comparing DIF in dichotomous variables, the difference in difficulty between two groups should be at least 0.5 measurement units with a p value <0.05 for DIF to be detected. When comparing more than two groups, the χ^2 statistic and p value <0.05 is used [18].

Floor and Ceiling Effects

The EAT-10 was also investigated for floor or ceiling effects. Floor or ceiling effects are considered to be present if more than 15 % of respondents achieved the lowest or highest possible score, respectively [28]. The presence of floor and ceiling effects are indicative that extreme items are missing in the lower or upper end of the scale, suggesting limited content validity.

Table 1 Description of the sample

Country	*N*	%
Spain	381	59.9
Italy	82	12.9
Turkey	90	14.2
Sweden	83	13.1
Total	636	100.0
Confirmation of diagnosis		
OD confirmed using gold standard	466	73.3
No-OD confirmed using gold standard	170	26.7
Total	636	100.0
Diagnoses		
Cardio vascular accident (CVA)	412	64.8
Neuro-degenerative disorder	90	14.2
Other diagnoses	60	9.4
Elderly	55	8.6
Head and neck cancer	17	2.7
Unknown	2	0.3
Total	636	100

Results

The multisite sample of 636 records from a clinical population at risk of having OD were used to conduct the analysis; 53.8 % were male and 46.2 % female and the mean age was 69.9 years (SD ± 13.9). Data were missing for only 27 item scores overall (<0.1 %). Data were drawn from four countries, and clinical diagnoses and outcome of gold standard assessment are presented in Table 1.

Rating Scale Validity

The EAT-10 is a 5-point rating scale and examination revealed that the as the category order increased (i.e., 0 through to 4), so did the average measure scores increase monotonically, and all were within an acceptable range resulting in five distinct and correctly ordered categories (see Table 2). Goodness of fit statistics were all acceptable (MnSq = 0.7–1.4) showing acceptable fit to the model; however, the majority of the scores (60 %) were 0 (no problem) which suggests that there were not enough difficult items. Examination of the Rasch–Andrich thresholds (see Table 2) revealed disordered thresholds with the threshold for categories 0–1 (none–5.58) and 1–2 (5.58 to −7.32) advancing by >5 logits and the threshold for category 3–4 advancing by <1.0. These findings would indicate that consideration should be given to collapsing categories 0 and 1 and of categories 3 and 4, where after rating scale use should be reexamined.

Person and Item Fit Statistics

The summary fit statistics for item and person ability demonstrated good fit to the model based on both *infit* and *outfit* statistics with a good reliability estimate (0.98) for items and are presented in Table 3. The person reliability measure was low (0.55) with a low person separation index of 1.11 rather than the required minimum of two levels. This indicates that the persons were not reliably separated into distinct groups based on a strata of ability. Principal component analysis of residuals revealed 264 (42 %) of people had mis-fitting MnSq *outfit* scores ($n = 91 > 1.4$; $n = 173 < 0.7$) indicating problems with internal consistency.

Item fit statistics are provided in Table 4. All items had an *infit* MnSq in the desired range (0.07–1.4); however, mis-fitting *infit* Z-STD scores > +2 for item 1 (lose weight), item 4 (solids effort), item 6 (painful), item 7 (pleasure eating (negative direction), and item 9 (cough) are of particular concern because they indicate more variation than modeled. Item 9 (cough) was the only outfit MnSq (1.48) that was mis-fitting, but the *outfit* standardized (*t*-statistic) for item 4 (solids effort–negative direction), item 7 (pleasure eat, negative direction), item 9 (cough), and item 10 (stressful–negative direction); all had Z-STD scores >2 (which is of greater concern than fit <0.07) and reflect more variation than modeled. *Infit* statistics provide more insight into how the items perform because *outfit* statistics are sensitive to outlying scores [25]. The negative directions (items 4, 7, and 10) indicate that the items do not contribute to the overall construct.

Table 2 Category order

Category	*N*	%	Average measures	Infit MnSq	Outfit MnSq	Andrich thresholds
0	3787	60	−11.12	1.07	1.11	None
1	551	9	−6.70	1.01	1.02	5.58
2	715	11	−3.22	0.93	0.77	−7.32
3	561	9	0.59	1.00	0.97	1.23
4	719	11	4.90	0.98	1.01	0.51

Table 3 Item and person summary statistics

	Reliability	Separation	Mean	Model	Infit		Outfit	
			Measure	SE	MnSq	Z-STD	MnSq	Z-STD
Item	0.98	7.50	50.00	0.43	1.03	0.1	0.98	−0.4
Person	0.55	1.11	38.48	6.70	0.98	0.0	0.99	0.1

Table 4 Individual item fit statistics and principal component analysis

#	Item	Mean	Model	Infit		Outfit		PCA
		Measure	SE	MnSq	Z-STD	MnSq	Z-STD	Factor loading
1	Lose weight	51.83	0.43	1.21	3.0	1.12	1.1	−0.40
2	Go out meals	49.16	0.41	0.99	−0.1	0.84	−1.7	−0.51
3	Liquids effort	47.86	0.40	0.99	−0.1	0.95	−0.5	−0.46
4	Solids effort	46.83	0.40	0.77	−4.1	0.72	−3.5	0.04
5	Pills effort	48.79	0.41	1.06	1.0	1.13	1.4	−0.19
6	Painful	58.96	0.59	1.26	2.4	1.19	1.1	0.53
7	Pleasure eat	49.21	0.41	0.84	−2.8	0.72	−3.1	−0.11
8	Stick throat	48.19	0.40	0.90	−1.7	0.91	−1.0	0.68
9	Cough	46.99	0.40	1.25	3.8	1.48	4.8	0.14
10	Stressful	52.18	0.44	0.98	−0.2	0.77	−2.1	0.42

PCA principal component analysis

Dimensionality

The Rasch dimension explained 48.3 % of the variance in the data, and >40 % is considered a strong measurement of dimension [18]. However, there was more unexplained variance (51.7 %) than explained. Of the 48.3 % explained variance, the item measures (34.3 %) explain more of the variance than the person measures (14.1 %). The raw variance explained by the items was more than four times the variance explained by the first contrast (8.4 %). The total raw unexplained variance (51.7 %) had an eigenvalue of 10, but the eigenvalue of first contrast was 1.62 which is less than the value (2 units) that would be required to indicate a second dimension. This indicates that the unexplained variance is too random to form a second dimension, but rather indicates that there are a number of items that do not contribute to the model (see Table 5).

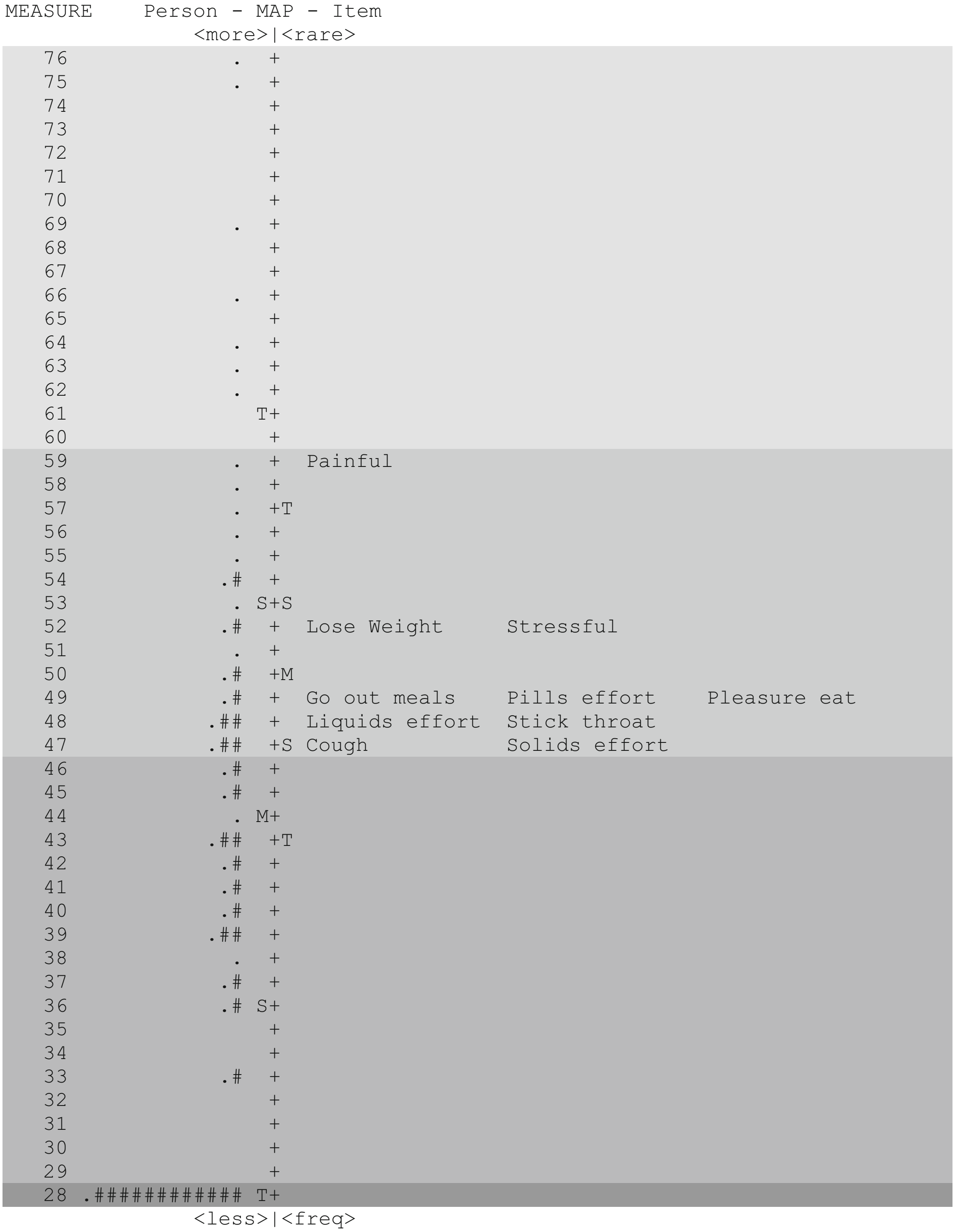
The principal component analysis (PCA) in Table 4 indicates that the items: swallowing solids (4) and pills (5) are an effort, pleasure to eat is affected (7), and cough when eating (9) showed very low loading. As presented in Fig. 1, the person–item dimensionality map shows that there are not enough easy and difficult items and that most people are not aligned with the items. Furthermore item redundancy is evident with several items aligning at the same level of person ability. For example, as can be seen in Fig. 1, many items are on the same level of difficulty for the items: go out meals, pills effort, and pleasure eat.

Differential Item Functioning (DIF)

The DIF analysis enabled examination of potential contrasting item-by-item profiles associated with: having or not having a diagnosis of OD; language; gender; and

Table 5 Standardized residual variance

Variance	Eigenvalue	Observed (%)	Expected (%)
Total raw variance in observations	19.35	100.0	100.0
Raw variance explained by measures	9.35	48.3	48.8
Raw variance explained by persons	2.72	14.1	14.2
Raw variance explained by items	6.63	34.3	34.6
Raw unexplained variance (total)	10.00	51.7	51.2
Unexplained variance in 1st contrast	1.62	8.4	16.2



```
MEASURE    Person - MAP - Item
               <more>|<rare>
   76             .  +
   75             .  +
   74                +
   73                +
   72                +
   71                +
   70                +
   69             .  +
   68                +
   67                +
   66             .  +
   65                +
   64             .  +
   63             .  +
   62             .  +
   61               T+
   60                +
   59             .  +  Painful
   58             .  +
   57             .  +T
   56             .  +
   55             .  +
   54            .#  +
   53             . S+S
   52            .#  +  Lose Weight     Stressful
   51             .  +
   50            .#  +M
   49            .#  +  Go out meals    Pills effort    Pleasure eat
   48           .##  +  Liquids effort  Stick throat
   47           .##  +S Cough           Solids effort
   46            .#  +
   45            .#  +
   44             . M+
   43           .##  +T
   42            .#  +
   41            .#  +
   40            .#  +
   39           .##  +
   38             .  +
   37            .#  +
   36            .# S+
   35                +
   34                +
   33            .#  +
   32                +
   31                +
   30                +
   29                +
   28 .############ T+
               <less>|<freq>
```

Fig. 1 Person–item dimensionality map. *Note* EACH "#" = 14: EACH "." = 1 to 13

diagnoses (other than OD). The summary of the DIF analysis is presented in Table 6 and revealed significantly different responses on items 1, 7, and 10 based on OD versus no-OD; items 2, 3, 6, 7, 9, and 10 based on language; items 2, 5, 6, and 7 on gender, and items 1, 3, 9, and 10 on diagnoses (other than OD). These results indicate that there was item bias which meant that the hierarchy of the items varied across samples.

Table 6 Summary DIF analysis

Items	Gender		OD vs. No-OD		Diagnostic categories		Language	
	DIF contrast	Mantel–Haenszel probability	DIF contrast	Mantel–Haenszel probability	Summary DIF χ^2	Probability	Summary DIF χ^2	Probability
1. Lose weight	−1.22	0.3445	−4.64	0.0216*	0.0216	0.0144*	5.450	0.1408
2. Go out meals	−1.73	0.0047*	−3.15	0.1654	0.1654	0.2202	10.670	0.0135*
3. Liquids effort	0.00	0.5022	−2.26	0.1540	0.1540	0.0026*	13.118	0.0043*
4. Solids effort	0.00	0.5728	−1.02	0.5347	0.5347	0.7005	5.952	0.1132
5. Pills effort	1.50	0.0312*	0.80	0.4638	0.4638	0.7925	3.983	0.2621
6. Painful	4.06	0.0264*	2.02	0.7918	0.7918	0.6244	11.550	0.0090*
7. Pleasure eat	−2.30	0.0085*	1.76	0.0305*	0.0305	0.7707	15.822	0.0012*
8. Stick throat	0.71	0.6410	1.35	0.1620	0.1620	0.7721	4.104	0.2493
9. Cough	0.00	0.8646	0.98	0.3920	0.3920	0.0044*	26.577	<0.0001*
10. Stressful	1.10	0.1551	3.66	0.0035*	0.0035	0.0414*	47.174	<0.0001*

$DF = 1$

* $p < 0.05$

Floor and Ceiling Effect

As there are 10 items with a minimum possible score of 0, and a maximum possible score of 4, the lowest possible sum of scores is 0 and the highest possible sum of scores is 40. Nearly 23 % ($n = 146$) of the 636 participants achieved the lowest possible score and 0.47 % ($n = 3$) achieved the highest possible score of 40. The findings demonstrate that more than 15 % of respondents achieved the lowest possible scores therefore indicating that the EAT-10 has floor effects (i.e., too many respondents achieved a total score of 0), but not ceiling effects. Figure 2 shows the spread of total item scores.

Discussion

We set out to evaluate the psychometric properties of the EAT-10 using Rasch analysis (IRT). The overall item reliability of the EAT-10 is good, and the overall item and person infit and outfit statistics were within acceptable parameters. However, the person reliability, an IRT equivalent for internal consistency (Cronbach's alpha), was poor. Furthermore, the person separation index was below the required parameter (<2), indicating the EAT-10 performs poorly in separating patients with different levels of swallowing problems into distinct groups.

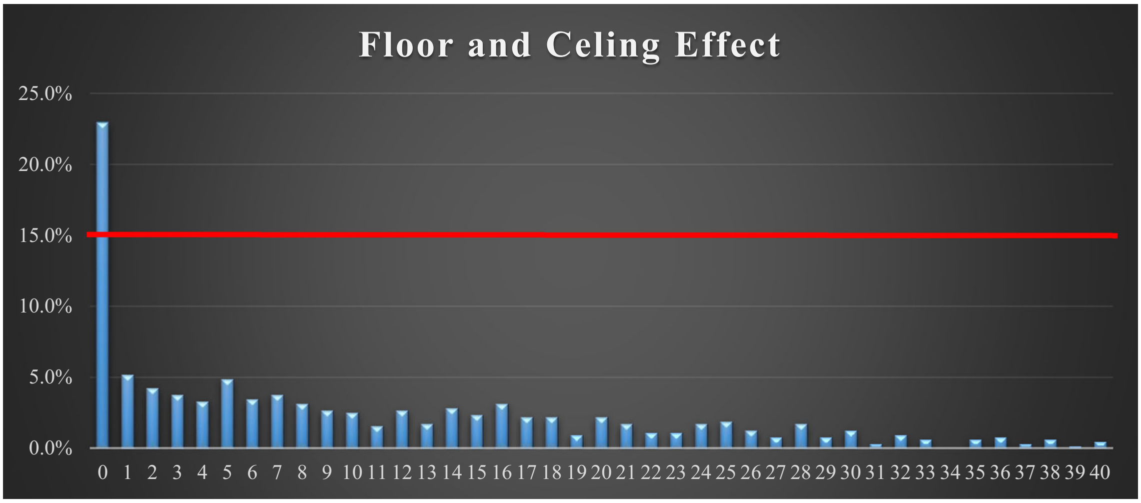


Fig. 2 Floor and ceiling effect

At an individual item level, the MnSq fit statistics were mostly within parameters with the exception of cough when eating (9); however, the Z-STD fit statistics for six of the items were outside the expected parameters. The dimensionality parameters indicated that the EAT-10 does not have a second dimension (thus is unidimensional), as such the unexplained variance is indicative of items not contributing toward the overall construct, rather than multidimensionality. Furthermore, a principal component analysis, equivalent to factor analysis, revealed that four items showed very low loading, which is indicative they do not contribute to the overall construct. Collectively on an item level, these findings indicate that several items are mis-fitting and that at least four of the ten items do not contribute toward the overall construct being measured.

An evaluation of the dimensionality of the measure using Rash analysis is indicative of structural validity. The large percentage unexplained variance (51.7 %) supports the finding that there are a number of items that do not contribute toward the overall construct and are indicative of poor structural validity. This finding is substantiated with the item DIF analysis where a number of items showed DIF against gender, having a confirmed diagnosis of OD or not, and patient diagnostic categories.

A DIF analysis was conducted of language (Italian, Swedish, Spanish, and Turkish) as a preliminary investigation of the EAT-10's cross-cultural validity. The finding that six of the ten items showed DIF on language is indicative that there may be problems with the EAT-10 translation into different languages. This warrants a more detailed investigation of cross-cultural validity, to examine if the exact procedures were followed during the translation process and if the translations meet international standards of cross-cultural validation, such as the guidelines stipulated in the Consensus-based standards for the selection of health measurement instruments (COSMIN) taxonomy [22].

Since the publication of the EAT-10 by Belafsky et al. [19], the EAT-10 has been translated in many different languages, including for example Spanish, Portuguese, Italian, and Arabic [20, 29–31]. Most studies referring to the reliability and validity of the EAT-10 concluded that the tool was both reliable and valid [for example, 31, 32, 20] and had excellent psychometric properties [1, 21]. The EAT-10 was one of the included tools in a psychometric review on functional health status questionnaires in OD [13]. Based on COSMIN taxonomy of measurement properties and definitions for health-related patient-reported outcomes [33], the EAT-10 obtained poor overall methodological quality scores for all measurement properties except for structural validity; as no data were reported in the literature, this psychometric property could not be rated. The authors concluded that the use of the EAT-10 in daily clinic and research could only be justified if the overall methodological quality would show satisfactory improvement on most measurement properties after re-evaluation.

The conclusions by Speyer et al. [13] seem to be in contrast with most of the publications on the reliability and validity of the EAT-10. Even though the review by Speyer et al. [13] included journal articles up to June 2013 and therefore did not cover the most recent literature, more prominent difficulties seem to be underlying to these identified discrepancies in psychometric assessment; first, authors use different psychometric concepts and definitions for psychometric properties. The COSMIN taxonomy, however, was developed through internationally expert consensus and has been utilized in an increasing number of psychometric publications. As a standardized tool, the COSMIN checklist [33], is used to evaluate the methodological quality of studies on any of the nine domains or psychometric properties. The COSMIN taxonomy describes relationships and definitions of psychometric properties and can be completed by quality criteria for measurement properties as defined by Terwee et al. [28] and Schellingerhout et al. [34]. In the literature, however, psychometric terminology is used in confusing ways. For example, criterion validity refers to the extent to which scores on a particular questionnaire relate to a gold standard [22, 28]. Santos Nogueira et al. [32] compared EAT-10 scores with a generic quality of life measure, the EQ-5D, and considered this criterion validity. However, the EAT-10 is a combination of items on functional health status and health-related quality of life [35]. Comparing the EAT-10 with a generic quality of life measures would seem hypothesis testing or convergent validity according to the COSMIN framework: the degree to which the scores on a particular questionnaire are consistent with hypotheses, for example, with regard to relationships to scores of other instruments.

Next, authors tend to generalize their findings; many authors stating that the EAT-10 is a reliable and valid tool, only considered a limited number of psychometric properties [for example, 20, 21, 32, 1, 31]. Most frequently, authors considered repeated measurements as part of reliability (but not intra and inter rater reliability), internal consistency, and convergent validity. In general, no reference was made to quality criteria for the assessment of the psychometric properties.

Classical Testing Theory (CTT) Versus Item Response Theory (IRT)

The most common frameworks used for developing measures and evaluating measurement properties are CTT and the more recently developed IRT [23]. While some areas of research have readily taken to IRT (e.g., psychology and education), the field of psychometric testing in allied health, especially in speech pathology, seems to continue using CTT. To our knowledge, the reliability and validity of the EAT-10 has not been evaluated using IRT.

CTT refers to a theoretical framework about test scores introducing the following concepts: test or observed score, true score, and error score. The assumptions in CTT are that true scores and error scores are uncorrelated, the average error score in the population of examinees is zero, and error scores on parallel tests are uncorrelated. The fact that CTT is based on these relatively weak assumptions and therefore easily met with test data is considered an advantage [23]. Even so, a number of limitations of CTT have been addressed in the literature; first, item difficulty and item discrimination are fundamental concepts within many CTT analyses but are group dependent. Another limitation of CTT is that scores are entirely test dependent, and as a consequence, test difficulty directly affects the resultant test scores. As such, CTT may be described as 'test based,' whereas IRT may be described as 'item based' [23]. Moreover, despite the added complexity, IRT has many advantages over CTT; IRT models estimate both item and person parameters with the same model. IRT determines person-free item parameter estimation and item-free trait level estimation. Further, IRT provides optimal scaling of individual differences and assists in analyses such as the evaluation of differential item functioning (DIF) [36]. As a consequence, CTT may lead to incorrect substantive conclusions, whereas IRT may retrieve more valid substantive findings [36].

Describing CTT and IRT in further detail is outside of the scope of this manuscript. However, taking into account some of the shortcomings of CTT and the potential benefits of IRT, psychometric research in allied health should not just place emphasis on CTT, but embrace the more recently developed IRT framework as well.

Conclusions

Using Rasch analysis to explore the psychometric qualities of the EAT-10 has exposed a number of difficulties. The EAT-10 has poor internal consistency and performs poorly in separating patients with different levels of swallowing difficulty into distinct groups. A combination of individual item misfit, high unexplained variance of the Rasch model and poor loading of items as reflected in the PCA analysis, demonstrated that the EAT-10 has a number of items not contributing toward the overall construct. Together with problems in the step calibration of the rating scale, the combination of findings is indicative of poor structural validity.

The person–item dimensionality map together with an evident floor effect indicates a combination of item redundancy, as well as the need for including both easier and more difficult items to better align person ability with item difficulty. A major flaw in the design of the measure is lack of item descriptors for the item categories 1, 2, and 3, which likely contributed toward respondents interpreting the item categories differently, resulting in a disordered use of the item categories. The paper highlights that the EAT-10 has significant problems in both reliability and validity. As such the EAT-10 should be redeveloped using IRT prior to further clinical use, given that the weaknesses of the EAT-10 is exposed on an item level (IRT), rather than on a whole test level (CTT).

Acknowledgments The authors express their gratitude to Cally Smith, Rebekah Totino, and Lauren Parsons for providing research assistant support.

Compliance with Ethical Standards

Conflict of interest The authors report no conflict of interest.

References

1. Rofes L, Arreola V, Mukherjee R, Clavé P. Sensitivity and specificity of the Eating Assessment Tool and the Volume–Viscosity Swallow Test for clinical evaluation of oropharyngeal dysphagia. Neurogastroenterol Motil. 2014;26(9):1256–65.
2. Sharma J, Fletcher S, Vassallo M, Ross I. What influences outcome of stroke—pyrexia or dysphagia? Int J Clin Pract. 2000;55(1):17–20.
3. Timmerman AA, Speyer R, Heijnen BJ, Klijn-Zwijnenberg IR. Psychometric characteristics of health-related quality-of-life questionnaires in oropharyngeal dysphagia. Dysphagia. 2014;29(2):183–98. doi:10.1007/s00455-013-9511-8.
4. Ekberg O, Hamdy S, Woisard V, Wuttge-Hannig A, Ortega P. Social and psychological burden of dysphagia: its impact on diagnosis and treatment. Dysphagia. 2002;17(2):139–46.
5. Stringer S. Managing dysphagia in palliative care. Prof Nurse. 1999;14(7):489–92.
6. Kertscher B, Speyer R, Fong E, Georgiou AM, Smith M. Prevalence of oropharyngeal dysphagia in the Netherlands: a telephone survey. Dysphagia. 2015;30(2):114–20.
7. Takizawa C, Kenworthy J, Gemmell E, Speyer R (under review) Oropharyngeal dysphagia in stroke, Parkinson's disease, head injury, community acquired pneumonia, and Alzheimer's disease: a systematic review. Cerebrovascular Diseases.
8. Perry L, Love CP. Screening for dysphagia and aspiration in acute stroke: a systematic review. Dysphagia. 2001;16(1):7–18.
9. Speyer R. Oropharyngeal dysphagia: screening and assessment. Otolaryngol Clin N Am. 2013;46(6):989–1008.
10. Cordier R, Speyer R, Chen Y, Wiles-Gillan S, Brown T, Bourke-Taylor H, Doma K, Leicht A. Evaluating the psychometric quality of social skills measures: a systematic review. PLoS ONE. 2015;10(7):e0132299.
11. Vaz S, Cordier R, Falkmer M, Ciccarelli M, Parsons P, McAuliffe T, Falkmer T. Should schools expect poor physical and mental health, social adjustment, and participation outcomes in

students with disability? PLoS ONE. 2015;10(5):e0126630. doi:10.1371/journal.pone.0126630.

12. Liamputtong P. Research methods in health: foundation for evidence-based practice. 2nd ed. South Melbourne: Oxford University Press; 2013.
13. Speyer R, Cordier R, Kertscher B, Heijnen BJ. Psychometric properties of questionnaires on functional health status in oropharyngeal dysphagia: a systematic literature review. BioMed Res Int. 2014. doi:10.1155/2014/458678.
14. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. J Nurs Scholarsh. 2005;37(4):336–42.
15. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. J Am Med Assoc. 1995;273(1):59–65.
16. Higginson I, Carr A. Using quality of life measures in the clinical setting. Br Med J. 2001;322(7297):1297–300.
17. Centers for disease control and prevention. Health-related quality of life (HRQOL). 2011. http://www.cdc.gov/hrqol/index.htm. Accessed Oct 2015.
18. Linacre JM. A user's guide to Winsteps® Rasch-model computer programs: program manual 3.92.0. Chicago: Mesa-Press; 2016.
19. Belafsky PC, Mouadeb DA, Rees CJ, Pryor JC, Postma GN, Allen J, Leonard RJ. Validity and reliability of the Eating Assessment Tool (EAT-10). Ann Otol Rhinol Laryngol. 2008;117(12):919–24.
20. Farahat M, Mesallam TA. Validation and cultural adaptation of the Arabic version of the Eating Assessment tool (EAT-10). Folia Phoniatr Logop. 2016;67(5):231–7.
21. Giraldo-Cadavid LF, Gutiérrez-Achury AM, Ruales-Suárez K, Rengifo-Varona ML, Barros C, Posada A, Romero C, Galvis AM. Validation of the Spanish version of the Eating Assessment Tool-10 (EAT-10spa) in Columbia. A blinded prospective cohort study. Dysphagia. 2015;. doi:10.1007/s00455-016-9690-1.
22. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. 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.
23. Hambleton KH, Jones RW. An NCME instructional module on: comparison of classical test theory and item response theory and their applications to test development. Educ Meas. 1993;12(3):38–47.
24. Linacre JM. WINSTEPS Rasch measurement computer program: version 3.92.0. Chicago: Winsteps; 2016.
25. Bond TG, Fox CM. Applying the Rasch model: fundamental measurment in the human sciences. 3rd ed. New York: Taylor & Francis; 2015.
26. Linacre JM. Investigating rating scale category utility. J Outcome Meas. 1999;3(2):103–22.
27. Smith EVJ. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In: Smith EVJ, editor. Introduction to Rasch measurement. Maple grove: JAM press; 2004. p. 93–122.
28. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HCW. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.
29. Burgos R, Sarto B, Segurola H, Romagosa A, Puiggrós C, Vázquez C, Cárdenas G, Barcons N, Araujo K, Pérez-Portabella C. Translation and validation of the Spanish version of the EAT-10 (Eating Assessment Tool-10) for the screening of dysphagia. Nutr Hosp. 2012;27(6):2048–55.
30. Rebelo Gonçalves MI, Bogossian Remaili C, Behlau M. Cross-cultural adaptation of the Brazilian version of the Eating Assessment tool-EAT-10. CoDAS. 2013;25(6):601–4.
31. Schindler A, Mozzanica F, Monzani A, Ceriani E, Atac M, Jukic-Peladic N, Venturini C, Orlandoni P. Reliability and validity of the Italian Eating Assessment Tool. Ann Otol Rhinol Laryngol. 2013;122(11):717–24.
32. Santos Nogueira D, Lopes Ferreira P, Azevedo Reis E, Sousa Lopes I. Measuring outcomes for dysphagia: validity and reliability of the European Portuguese Eating Assessment Tool (P-EAT-10). Dysphagia. 2015;30:511–20.
33. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, De Vet HC. 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.
34. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res. 2012;21(4):659–70.
35. Speyer R, Kertscher B, Cordier R. Functional health status in oropharyngeal dysphagia. J Gastroenterol Hepatol Res. 2014;3(5):1043–8.
36. Reise SP, Henson JM. A discussion of modern versus traditional psychometrics as applied to personality assessment scales. J Personal Assess. 2003;81(2):93–103.

R. Cordier PhD

A. Joosten PhD

P. Clavé PhD

A. Schindler PhD

M. Bülow PhD

N. Demir PhD

S. Serel Arslan PhD

R. Speyer PhD