Using Rasch analysis to evaluate the reliability and validity of the Swallowing Quality of Life questionnaire: an item response theory approach

[1]Cordier, R., [2]Speyer, R., [3]Schindler, A., [4]Michou, E., [5]Heijnen, B.J., [6]Baijens, L.W.J., [7]Karaduman, A., [8]Swan, K., [9]Clave, P., & [1]Joosten, A.

[1]Curtin University Australia;

[2]James Cook University Australia;

[3]University of Milan Italy;

[4]University of Manchester United Kingdom;

[5]Leiden University Medical Center Netherlands;

[6]Maastricht University Medical Center Netherlands;

[7]Hacettepe University Turkey

[8]Gold Coast Health Service, Queensland Health, Australia

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

Corresponding Author:

A/Prof Reinie Cordier, School of Occupational Therapy and Social Work, Curtin University, Perth, WA, Australia. Email: r.cordier@curtin.edu.au

Using Rasch analysis to evaluate the reliability and validity of the Swallowing Quality of Life questionnaire: an item response theory approach

Abstract

Introduction: The Swallowing Quality of Life questionnaire (SWAL-QOL) is a 44-item questionnaire that consists of 11 subscales and is widely used both clinically and in research to evaluate patients' quality of life related to swallowing difficulties. The SWAL-QOL has been described in literature as a valid and reliable tool with alpha values for internal consistency ranging between 0.79 and 0.91 for the different subscales, test-retest reliability ranged between 0.60 and 0.91 and ICC ranged from 0.59 to 0.91. However, the SWAL-QOL was developed and tested using classic-test theory. This study describes the reliability and validity of the SWAL-QOL using item response theory (IRT; Rasch analysis).

Methods: SWAL-QOL data were gathered from 4 European countries involving 507 participants (MN age = 63.7 years; SD = 12.8), all at risk of oropharyngeal dysphagia (OD). Most patients (83%) underwent videofluoroscopy and/or Fiberoptic endoscopic evaluation of swallowing to confirm OD; the remaining 17% received a clinical diagnosis based on meeting selected clinical criteria; 75.7% OD and 24.3% no OD. To ensure the sample was homogenous, patients with esophageal dysphagia were excluded. Data was analysed using Rasch analysis.

Results: When analysing all the items combined, the overall item and person reliability of the SWAL-QOL was good, 0.98 and 0.94 respectively. However, the person reliability was poor for 8 of the 11 subscales (0.47-0.73) and the item reliability was poor for the fear subscale. Eight of the subscales exhibited poor person separation and two subscales exhibited poor item separation. The overall item and person fit statistics were within acceptable range (< 1.4 MnSq; > 2/ Z- <-2 Z-STD). However, on an individual item fit level, twenty-eight items had infit values outside the acceptable range, indicating unpredictable responses for these items and 10 items had large negative outfit values indicating item redundancy. The item-person dimensionality map confirmed this finding and also demonstrated that a large proportion of the items is on the

same difficulty level. The overall Rasch model fit demonstrated high unexplained variance (59.5%); while Principal Component Analysis showed high unexplained variance in the first contrast, suggesting a second dimension. For all the items combined, none of the item categories were 'category', 'threshold' or 'step' disordered; however, on a subscale level all subscales demonstrated some form of category disordered functioning.

Conclusions: The findings suggest an urgent need to further investigate the underlying structure of the SWAL-QOL and its psychometric characteristics using IRT.

Background

The ability to swallow is impaired in a range of common conditions such as stroke, head and neck cancer and neurological diseases [1, 2]. Disordered swallowing is known as oropharyngeal dysphagia (OD); prevalence varies depending on aetiology and definition. OD affects 8.1–80% of stroke patients, 11–81% of people with Parkinson's disease and up to 30% of traumatic brain injury patients [3]. Additionally, OD may be present in 13-57% of patients with dementia [4], and a 2015 study by Kertscher, Speyer [5] reported the prevalence ranged from 2.3-16% among the general population with the incidence rising with advanced age. As population life expectancy and diseases related to aging increase, it may be expected OD prevalence will likewise increase [1].

OD affects physical wellbeing, and may result in dehydration, aspiration pneumonia or even death [6]. In order for clinicians to effectively manage OD it is important to consider a patient's functional health status (FHS), and evaluate the impact of OD on the functional aspects of patient health, such as the ability to perform various eating and swallowing tasks [7, 8]. However, given that OD is associated with aging and chronic diseases [4, 5] and individuals may live with the condition for many years, the long-term impact of OD on psychosocial well-being is also an important consideration in the effective management of OD. Quality of life encompasses psychosocial wellbeing as well as mental and physical health [9]. The impact of a disease on quality of life is known as Health-Related Quality of Life (HR-QOL) [7]. The HR-QOL of people with OD may be affected by feeling isolated at mealtimes, being unable to enjoy meals or feeling anxious and distressed by the everyday act of eating and drinking [10-12].

As OD may have significant effects on HR-QOL, it is important to measure HR-QOL when evaluating the efficacy of OD interventions. A range of assessments have been used to investigate HR-QOL in OD; they are typically self-assessment questionnaires [13, 14] and several are specific to certain diseases, such as the Radboud Oral Motor Inventory for Parkinson's Disease [15]. Assessments which target the generic OD population include the Swallowing–Quality of Life Questionnaire (SWAL-QOL) [16], Dysphagia Handicap Index (DHI)

[17] and Deglutition Handicap Index (DHI) [18]. Although all these measures examine HR-QOL, they differ in terms of the domains of HR-QOL being assessed, the number of items, response options, scales and scoring systems [14].

Measures need to have sound psychometric properties to accurately report change in the phenomenon under study [9]. Psychometric properties refer to quantifiable data that describe the statistical strengths and weaknesses of an assessment, such as validity and reliability [19]. Two reviews have examined the psychometric quality of HR-QOL measures for OD. Timmerman, Speyer [14] evaluated the psychometric properties of HR-QOL measures using the quality criteria for measurement properties of health-status questionnaires recommended by Terwee, Bot [20]. The authors awarded high scores on criteria for face validity, criterion validity and interpretability; prerequisites for appropriate use of HR-QOL measures. Among the measures analysed, the SWAL-QOL achieved the highest scores. Keage, Delatycki [13] reviewed self-reported assessments of OD for patients with neurological diseases. Many of these measures included HR-QOL assessment as well as subscales reporting on swallow function. Reliability, validity, interpretability, responsiveness and precision of the measures were assessed using the criteria outlined by Fitzpatrick, Davey [21]. The authors likewise reported that the SWAL-QOL showed the strongest combination of psychometric properties in terms of reliability and convergent validity.

The SWAL-QOL assesses the patient's perspective on their swallowing. It is a 44-item measure on HR-QOL and consists of ten short (2-5 item) subscales as follows: burden of dysphagia, food selection, eating duration, eating desire, fear related to eating, sleep habits, fatigue, communication difficulties, mental health, social functioning [16]. There is also an additional 14-item subscale on frequency of dysphagia symptoms [16]. The SWAL-QOL's alpha values for internal consistency range between 0.79 and 0.91 for the different subscales, test-retest reliability range between 0.60 and 0.91 and intra-class correlation (ICC) range between 0.59 to 0.91 [16].

Although the literature to date reports the SWAL-QOL is the most psychometrically sound of HR-QOL measures for OD available, the SWAL-QOL has only been examined using Classic Testing Theory (CTT) [14]. CTT and the more recently developed Item Response Theory (IRT) are the most common frameworks used for developing measures and evaluating psychometric properties [22]. Even though procedures and interpretation of CTT are relatively straight forward compared with IRT, CTT has some limitations. The CTT framework evaluates the performance of the measure as a whole, rather than assessing the reliability of each item within the measure and its contribution to the overall construct. The evaluation is also specific to the sample population the measure was tested with. By contrast, in IRT the item is the unit of analysis and results are not bound by the test population [23].

The purpose of this study was to apply an IRT approach to investigate the reliability and validity of the SWAL-QOL. Using the Rasch measurement model, this study aimed to evaluate the response scale, the person and item fit characteristics, the dimensionality of the scale, and differential item function.

Methods

Participants

Five 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), Leiden University Medical Center and Maastricht University Medical Center (Netherlands), University of Milan (Italy) and University of Manchester (UK). To maximise homogeneity in the clinical swallowing characteristics of the sample, patients with confirmed esophageal dysphagia were excluded. Only those in the clinical population deemed at risk of OD, after initial intake (patient history and/or screening) were included.

Protocol

All patients completed the SWAL-QOL after which a videofluoroscopic or fiberoptic endoscopic evaluation of swallowing (VFS or FEES) recording of swallowing was performed as part of

standard clinical practice or usual care for 83% of patients and 17% met clinical diagnosis based on consensus assessment by a dysphagia team consisting of two speech therapists and a laryngologist. 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. Patient characteristics were collected on both gender and age.

The original version of the SWAL-QOL by McHorney, Robbins [16] was published in English (see Supplementary File of the full scale with complete item descriptors). This study used the original English version and translations of the SWAL-QOL into three different languages: Turkish, Dutch, and Italian. 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 trialled by pre-testing in patients at risk for OD to check the ease of comprehension, the interpretation and cultural relevance of the SWAL-QOL items.

Statistical Analysis

Data were analysed using Winsteps version 3.92.0 [24], with the joint maximum likelihood estimation rating scale estimation [25]. The reliability and validity of the SWAL-QOL was evaluated using Rasch analyses in a three stepped approach. First, Rasch analyses were performed for all (44) items of as a measure of quality of life. Second, Rasch analyses were performed on each of the eleven subscales that comprise the overall measure (i.e. the 10 short subscales and the 14-item symptom scales). Third, Rasch analyses were performed for two of the three factors described in McHorney, Robbins [16]. *Factor One* (Dysphagic-specific QOL) consisted of the items related to the following subscales: Food selection, Burden, Mental health, Social functioning, Fear, Eating duration, Eating desire, and Communication, while *Factor Two* (Generic QOL) consisted of the subscales Sleep and Fatigue. The factor on Quality of Care (clinical information, general advice, and patient satisfaction) was not analysed as this information was not available for participants. The following analyses were conducted for all three investigations.

Rating scale validity

The SWAL-QOL utilises a 5-point ordinal response scale for all items. Implicit in the use of ordinal scales is an assumption that higher ratings indicate "more" of the concept under assessment and the converse to be true for lower ratings. Examining rating scale validity can determine whether this is in fact the case for a particular scale. Rating scale response options are henceforth referred to as *categories*, and the categories are numbered 1-5 in alignment with the SWAL-QOL response options.

The SWAL-QOL's response scale was examined for both category and step (threshold) disordering. *Category disordering* was evaluated to determine if the rating response scales were being used in the expected manner, by examining rating categories for even distribution. To do this we examined whether average measure scores (frequency of use) increased monotonically as the category increased, which is indicative of ordered categories. Non-uniformity occurs when there are poorly defined categories or with the inclusion of items that do not measure the construct. *Category misfit*, indicated by Fit means squares (MnSq) outside 0.7 - 1.4, also shows category disordering and consideration should be given to collapsing it with an adjacent category [23].

Andrich-thresholds, or step calibrations (the point at which it is equally likely that the response is either of two adjacent categories), were examined to evaluate *step disordering*. Andrich-thresholds reflect the distance between the categories and should progress monotonically (i.e. there should be no overlap between categories, nor too large a gap between categories). On a 5-category scale the average measure distinct categories are indicated by an increase of at least 1.0 logit; however, an increase of >5.0 logits is indicative of gaps in the variable [26]. Step disordering does not indicate that the category definitions are out of sequence, rather that the category defines a narrow section of the variable.

Person and item fit statistics

To determine if the scale was a valid measure of the construct, fit statistics were used to identify misfitting items and the pattern of responses for each person. In this study, interpretation of fit

statistics, reported as log odd units (logits), indicate whether the items contribute to the one construct (i.e. the impact of swallowing disorders on everyday life) and the extent to which any one person's responses are reliable. Infit and outfit are reported as unstandardized MnSq or Z-Standard (Z-STD) scores. Infit and outfit fit reported as a MnSq should have a value close to 1.0 with an acceptable range of 0.7 - 1.4 for rating scales [27]. The expected outfit Z-STD values is 0 and values that exceed ±2 are interpreted as less than the expected fit to the model [27]. Model underfit degrades the model and requires further investigation to determine the reason for the underfit, while overfit could result in a misinterpretation that the model worked better than expected, but does not always degrade the model [27].

The person reliability is equivalent to the traditional Cronbach's alpha and is indicative of the measure's internal consistency. Low person reliability values (< 0.8) indicate a narrow range of person measures (i.e., not having enough persons with more extreme abilities, both high and low), or having too few items.

Person separation (if the outlying measures are accidental) and person separation index (PSI)/strata (if the outlying measures represent true performances; 4*person separation +1/3) are used to classify people. Person separation reports whether the test separates the sample into enough levels to determine high performers from low performers. Low person separation suggests the instrument is not sensitive enough to separate high and low performers. Reliability of 0.5 indicates separations into only one or two levels,0.8 indicates separation into 2-3 levels, and 0.9 indicating separation into 3 or 4 levels [23]. A PSI/strata of 3 is needed to consistently identify three different levels of performance (the minimum level required to attain a reliability of 0.9). Item reliability verifies item hierarchy with <3 levels (high, medium, low) with item reliability < 0.9 indicating the sample is too small to confirm the construct validity (item difficulty) of the instrument.

Dimensionality of the scale

Dimensionality is examined by: (a) finding any potentially problematic items evident by negative point-biserial correlations; (b) using Rasch fit indicators to identify misfitting persons

or items; and (c) by conducting Rasch factor analysis using Principal Component Analysis (PCA) of the standardised residuals [28]. PCA of residuals is used to check that there are no principal components (second or further dimensions) after the intended or Rasch dimension is removed. If the residuals for pairs of items are uncorrelated and normally distributed this indicates there is no second dimension. Recommended criteria for determining if there are further dimensions in the residuals: (a) A cut-off of >60% of the variance explained by the Rasch factor; (b) an eigenvalue of <3 (equivalent to three items) on first contrast, and (c) percentage variance explained by first contrast of <10% [23].

The distributions of the person abilities and item difficulties are schematically represented in the person–item dimensionality map using a logit scale. In the context of evaluating QOL, person ability refers to the level of QOL reported by a respondent, and item difficulty can be conceptualised such that "difficult" items evaluate an aspect of QOL that occurs with such rarity that very few responders will give a high rating to that item, and "easy" items evaluate an aspect of QOL that occurs relatively commonly such that all responders will give a high rating to that item. Where two or more items represent similar difficulty they occupy the same location on the logit scale. Locations on the logit scale where persons are represented with no corresponding item are gaps in the item difficulty continuum. The person measure score is another indicator of overall distribution. If the person measure score location is lower than the centralised item mean measure score (50), then it is indicative that the people in the sample were more able than the level of difficulty of the items. If the mean person location is higher (above 50), then the people in the sample was less able than the mean item difficulty.

Differential item analysis

In order to examine whether the scale items were used in the same way by all groups, a differential item analysis (DIF) was conducted. DIF occurs when a characteristic of the respondent other than their ability on the underlying trait influences their response to an item [27].

For DIF analysis, the sample was categorised by gender (male/female), language (Dutch, Italian, Turkish and English), medical diagnosis (Cardio Vascular Accident [CVA], Neuro-Degenerative Disorders, Elderly, Head and Neck Cancer, and other diagnoses), and the presence or absence of oropharyngeal dysphagia (OD vs. No OD). In determining DIF when comparing two groups (i.e., gender and OD vs No OD) with the hypothesis "this item has the same difficulty for two groups" the DIF contrast, which is the difference in difficulty of the item between the two groups, should be at least 0.5 logits with a p-value < 0.05 for DIF to be noticeable. In determining DIF when comparing more than two groups (i.e., medical diagnosis and language) with the hypothesis “this item has no overall DIF across all groups”, the chi-square statistic and p-value < 0.05 is used [23].

Results

The multi-site sample of 507 records from a clinical population with and without a confirmed diagnosis of oropharyngeal dysphagia (OD) were analysed; 59.2% were male and 40.8% were female and the mean age was 63.7 years (SD 12.8). Data were missing for 174 responses for 507 patients across 44 items, thus representing 0.78% of all possible response options. On item level, missing data ranged from 0% missing data per item to 2.96% missing data per item. Thus the possible influence of missing data was deemed negligible. Data were collected in four countries; the medical and demographic data of participants are presented in Table 1.

Table 1 Description of the sample

Country	***N***	**%**
Netherlands	296	58.4
Italy	87	17.2
Turkey	98	19.3
UK	26	5.1
Total	507	100
Confirmation of diagnosis	**N**	**%**
OD confirmed using Gold-standard	331	65.3
OD confirmed with clinical diagnosis	53	10.5
No OD confirmed using Gold-standard	90	17.8
No OD confirmed clinical diagnosis	33	6.5
Total	507	100
Clinical Diagnoses	**N**	**%**
Cardio Vascular Accident (CVA)	116	22.9
Neuro-Degenerative Disorder	234	46.2
Head and Neck Cancer	75	14.8
Other diagnoses	82	16.2
Total	507	100
Educational Status	**N**	**%**
Elementary school	65	12.8
High school 5 grades	33	6.5
High school 6 grades	121	23.8
Bachelor degree	71	14.0
Master degree	60	11.8
Doctorate degree	28	5.5
Unknown	130	25.6
Total	507	100
Marital status	**N**	**%**
Never married	40	7.9
Married	278	54.8
Divorced	19	3.8
Widowed	33	6.5
Unknown	137	27.0
Total	507	100

Rating Scale Validity

The SWAL-QOL is a 44-item measure that assesses eleven quality of life concepts (represented by eleven subscales) for people with OD and is reported to be sensitive to the clinically defined differences in dysphagia severity [16]. It uses a 5-point rating scale (1-5) which comprise the 5 categories analysed here. First we examined the overall instrument (all 44 items combined), followed by analysing the subscales individually, and finally we analysed the data using the two factors as reported in McHorney, Robbins [16] (*Factor One*: dysphagia specific QOL items and *Factor Two*: Generic QOL items; see Table 2). When examining the category order for the overall instrument (all 44 items combined), the average measures increased monotonically and all were in an acceptable fit range resulting in five distinct, ordered categories. All fit statistics were in the acceptable range (MnSq = .7 to 1.4). Examination of the Andrich thresholds (see Table 2) revealed disordered thresholds for categories 4-5 (decreasing from 3.74 to 1.30) and, although the remaining categories were not disordered, the advance between categories 3 and 4 was > 5 logits (5.18), indicating potential gaps in the variable.

We then examined category order of each subscale and the two factors, all average measure scores increased monotonically; however, examination of category fit statistics revealed some subscales in the misfit range as presented in Table 2. Infit and outfit MnSq for category 1 in Burden, and category 1 for Mental Health, and categories 3 and 4 for Food Selection, category 3 for Communication and category 5 for Eat Duration were outside the fit range. Infit for category 5 for Burden, categories 1 and 5 for Food Selection and Communication and outfit for category 3 on Eating Desire were outside the recommended range.

The magnitude of the distances between the thresholds exceeded the 5 logits limit for all adjacent categories for seven subscales and Factor Two (see Table 2), indicating gaps in the variable. All remaining subscales had at least one or more adjacent categories exceeding the 5 logits limit. The only exception is Factor One where none of the adjacent categories exceeded the limit.

Table 2 Category function

Scales	Category	N	%	Average measures	Infit MnSq	Outfit MnSq	Andrich thresholds
				All Items			
All Items	1	2836	13	-3.70	1.03	1.16	NONE
	2	3251	15	-.59	.98	.97	-3.60
	3	4062	18	2.34	.99	1.03	-1.44[c]
	4	4107	19	4.91	.96	.85	3.74[c]
	5	7834	35	10.18	1.01	1.06	1.30[d]
				Subscales			
Burden	1	189	19	-34.45	1.70[a]	1.63[a]	NONE
	2	175	17	-25.50	.87	.90	-43.48
	3	212	21	-1.35	.78	.76	-14.06
	4	196	19	26.76	.59	.59	13.40
	5	235	23	39.96	1.45[a]	1.37	44.14
Eating duration	1	273	27	-35.72	1.25	1.22	NONE
	2	155	15	-19.44	.93	.98	-28.94[c]
	3	161	16	.02	.70	.66	-9.67[c]
	4	147	15	18.05	.77	.74	8.99[c]
	5	269	27	35.06	1.52[a]	1.42[a]	29.61[c]
Eating desire	1	177	12	-8.59	1.24	1.18	NONE
	2	171	11	-4.00	.83	.76	-9.50[c]
	3	181	12	2.07	.87	.68[b]	-.56[c]
	4	227	15	9.48	.84	.97	3.08
	5	751	50	11.88	1.09	1.04	6.98
Symptoms	1	781	11	-4.65	1.10	1.14	NONE
	2	1113	16	-1.28	.96	.92	-7.05[c]
	3	1620	23	2.37	.97	.94	-3.12
	4	1148	16	6.05	.99	.92	7.86[c]
	5	2347	33	11.58	1.00	1.02	2.31[c]
Food Selection	1	96	10	-22.03	1.78[a]	1.38	NONE
	2	142	14	-16.78	.84	.83	-36.27[c]
	3	124	12	-.33	.60[b]	.54[b]	-5.44[c]
	4	292	29	20.88	.63[b]	.66[b]	1.76
	5	353	35	28.44	1.70[a]	1.16	39.94[c]
Communication	1	127	13	-24.59	1.43[a]	1.32	NONE
	2	162	16	-16.42	.84	.90	-31.11[c]
	3	177	18	1.11	.66[b]	.61[b]	-7.24[c]
	4	206	20	18.42	.72	.74	7.02[c]
	5	335	33	26.69	1.57[a]	1.38	31.34[c]
Fear	1	171	8	-8.79	1.14	1.17	NONE
	2	167	8	-4.06	.76	.78	-8.95[c]
	3	275	14	4.23	.99	.98	-4.46
	4	349	17	9.51	.74	.85	3.95[c]
	5	1057	52	14.08	1.17	1.14	9.46[c]
Mental health	1	364	14	-18.79	1.51[a]	1.48[a]	NONE
	2	343	14	-9.00	.86	.88	-17.98[c]
	3	469	19	2.45	.81	.78	-5.42[c]
	4	378	15	12.17	.80	.91	9.39
	5	973	39	20.20	1.05	1.07	14.01[c]

Scales	Category	N	%	Average measures	Infit MnSq	Outfit MnSq	Andrich thresholds
Social functioning	1	351	14	-22.13	1.05	1.10	NONE
	2	358	14	-11.12	1.00	1.03	-20.89[c]
	3	349	14	-.18	.80	.84	-5.09[c]
	4	583	23	12.40	.83	.84	.53
	5	878	35	25.68	1.33	1.21	25.44[c]
Fatigue	1	179	12	-27.60	1.07	1.05	NONE
	2	306	20	-13.15	.98	.98	-29.90[c]
	3	318	21	1.32	.81	.78	-5.89[c]
	4	387	26	17.26	.85	.86	7.15
	5	326	22	28.35	1.37	1.28	28.65[c]
Sleep	1	129	13	-20.67	1.57	1.37	NONE
	2	171	17	-13.44	.76	.78	-29.36[c]
	3	183	18	2.50	.77	.69	-4.79[c]
	4	202	20	16.63	.83	.84	7.78[c]
	5	326	32	23.40	1.27	1.20	26.38[c]
Factors							
Factor One	1	1748	14	-6.11	1.01	1.27	NONE
	2	1673	13	-1.68	1.00	1.00	-3.65
	3	1948	15	2.00	.95	1.03	-1.57
	4	2378	19	5.98	.87	.81	2.00
	5	4851	39	12.68	1.07	1.16	3.22
Factor Two	1	308	12	-14.81	1.04	1.04	NONE
	2	477	19	-6.26	.98	.97	-17.00[c]
	3	501	20	1.92	.86	.80	-2.56[c]
	4	589	23	9.99	.92	.92	4.13[c]
	5	652	26	17.76	1.18	1.17	15.43[c]

Notes. [a] Infit or Outfit MnSq >1.4; [b] Infit or Outfit MnSq <0.7; [c] Andrich threshold category increase of >5; [d] Andrich threshold category decrease where an increase is expected

Category probability curves

The category probability curves (Supplementary Figure 1) provide a visual means of examining the distinctions between thresholds to see if each response category had a distinct peak to indicate that each category was the most probable response for some portion of the variable [27]. As can be noted from the visual examination of the category probability curves, all items combined and the Symptom subscale were step disordered.

Item and Person Summary statistics

The summary fit statistics for item and person ability for all 44 items combined demonstrated good fit to the model based on both infit and outfit statistics with a good item reliability estimate (0.98). The person reliability was high (0.94) with a PSI of 5.51, which is higher than the required minimum PSI of 3 to reliably separate people into distinct strata of ability, and are presented in Table 3 along with the summary fit statistics for each subscale and the two factors. PCA of residuals revealed 214 (42.2%) of people had misfitting MnSq outfit scores (n = 101 >1.4; n = 113 < 0.7) indicating problems with internal consistency.

Examination of the summary fit statistics for each subscale revealed low person reliability for most subscales (range .47 - .73) resulting in PSIs for the subscales with low reliability in a range of <2 - <3. This means that for those subscales people are not being separated into at least two levels of ability. Low person reliability can indicate the need for more items in each subscale in order to separate high and low performers, or to introduce a broader sample of people ability [23]. The only exceptions were for the subscales Symptoms (.83), Mental health (.80), and Social functioning (.83). The person reliability scores for Factor one (.91) and Factor Two (.81) were within acceptable range. Item reliability for all scales was > .9 for All items, all the subscales and Factors, which confirms the hierarchy of the subscale items. The only exception was the subscale Fear which had a low item reliability (0.50).

Table 3 Item and person summary statistics

	Item/ Person			Person Separation	Mean	Model	Infit		Outfit	
Scales		**Reliability**	**Separation**	**Index**	**Measure**	**SE**	**MnSq**	**Z-STD**	**MnSq**	**Z-STD**
All Items	**Item**	.98	7.06	-	50.0	.41	1.01	-.1	1.04	.3
	Person	.94	3.88	5.51	54.34	1.53	1.04	-.1	1.04	-.1
Burden	**Item**	.97	5.31	-	50.0	.93	.99	-.1	.97	-.4
	Person	.71[a]	1.56[c]	2.41[d]	51.75	12.49	.92	-.3	.93	-.3
Eating duration	**Item**	1.00	17.07	-	50.0	.85	.99	-.1	.96	-.4
	Person	.66[a]	1.4[c]	2.2[d]	48.36	11.44	.85	-.3	.92	-.2
Eating desire	**Item**	.84	2.25	-	50.0	.57	1.01	.0	.93	-.8
	Person	.47 [a]	.95[c]	1.6 [d]	54.32	6.65	.93	.0	.93	.0
Symptoms	**Item**	.99	9.24	-	50.00	.44	1.02	.1	1.00	.0
	Person	.83	2.19	3.25	54.55	2.79	1.01	-.1	1.01	-.1
Food Selection	**Item**	.93	3.56	-	50.00	.89	.98	-.2	.86	-1.7
	Person	.53 [a]	1.06[c]	1.75 [d]	57.58	12.00	.85	-.4	.86	-.4
Communication	**Item**	.98	7.21	-	50.00	.81	.98	-.2	.95	-.7
	Person	.57 [a]	1.16[c]	1.88 [d]	53.30	10.85	.93	-.3	.94	-.3
Fear	**Item**	.50 [a]	.99[b]	-	50.00	.60	1.01	-.4	1.00	-.4
	Person	.56 [a]	1.13[c]	1.84[d]	56.57	6.05	.99	-.1	1.00	-.1
Mental health	**Item**	.98	6.25	-	50.00	.65	1.00	-.3	1.00	-.4
	Person	.80	2.02	3.03	54.01	6.02	.99	-.2	1.00	-.2
Social functioning	**Item**	.94	3.87	-	50.00	.72	.99	-.4	.99	-.4
	Person	.83	2.21	3.28	53.20	6.58	1.00	-.2	.99	-.2
Fatigue	**Item**	.98	7.59	-	50.00	.70	1.00	-.2	.98	-.5
	Person	.73 [a]	1.66[c]	2.55 [d]	54.48	8.82	.98	-.2	.98	-.2
Sleep	**Item**	.97	5.65	-	50.00	.75	.99	-.2	.94	-.8
	Person	.54 [a]	1.09[c]	1.78 [d]	54.11	10.28	.93	-.2	.94	-.2
Factor One	**Item**	.98	6.7	-	50.00	.44	1.02	-.3	1.09	.3
	Person	.91	3.19	4.58	54.93	2.30	1.05	-.1	1.09	.0
Factor Two	**Item**	.97	5.76	-	50.00	.56	1.01	.0	.99	-.3
	Person	.81	2.05	3.07	53.99	5.72	1.00	-.1	.99	-.1

Notes. [a] Person or item reliability <0.8; [b] Item separation <3.0; [c] Person separation <2.0; [d] Person separation index <3.0

Item Fit statistics

Point biserial correlations were examined to identify potentially misfitting items and to ensure all were in a positive direction indicating they potentially contribute to the overall construct. Item misfit was examined for all 44 items combined (see Table 4). Because they are unweighted, outfit statistics are often regarded as less important than infit statistics, but it is important to examine for contradiction between infit and outfit scores. As can be seen in Table 4, although there are more reported outfit Z-STD scores that fail to fit the model than infit Z-STD scores, there are no contradictions in the extent or direction of the misfit.

Underfit (i.e., too much variation as responses are too haphazard) is of greater concern than overfit as it can degrade the model [27]. With underfit (MnSq > 1.4; Z-STD > 2) being regarded as the bigger threat to the measure, this was examined first. Underfit of both infit and outfit scores (MnSq > 1.4; Z-STD > 2) was observed for the following items: Don't care, Drool, and Don't fall asleep. Overall more misfit was evident on both the infit and outfit Z-STD scores than the MnSq scores. Items that were more underfitting when the scale was used as a whole, but only on Z-STD scores (both infit and outfit), were Longer, Dribbling nose, Excess saliva, Thick saliva, Fear pneumonia and Stay asleep.

Overfit can result from item interdependence and caution needs to be taken in over reporting the quality of the test as the negative direction of these overfit scores indicates not enough variation (i.e., if all easy items are correct and then all difficult items are incorrect). Both MnSq and Z-STD infit and outfit scores were outside the model (overfit- MnSq < 0.7; Z-STD < -2) for the following items: Social life, Change work, and Social gathering and for infit scores on the item Dealing and for outfit on the items Discouraged and Role. Items that were more overfitting (Z-STD < -2) on both infit and outfit scores were Distracting, Choke food, Figure out, Annoyed, Frustrated, Impatient, and Not go out.

Table 4 Individual item fit statistics and principal component analysis for all 44 items combined

			Infit		Outfit		Factor	Point biserial
Items	**Measure**	**SE**	**MnSq**	**Z-STD**	**MnSq**	**Z-STD**	**loading**	**Correlations**
Dealing	54.01	.39	.67[b]	-6.7[c]	.79	-3.4[c]	32	.68
Distracting	52.21	.39	.85	-2.9[c]	.85	-2.3[c]	41	.65
Longer	57.48	.41	1.25	3.9[c]	1.29	3.8[c]	.08	.57
Forever	50.82	.40	1.05	.9	.99	-.1	.14	.58
Don't care	46.86	.43	1.43[a]	6.1[c]	1.42[a]	4.5[c]	.03	.44
No hunger	48.53	.41	1.34	5.2[c]	1.44[a]	5.0[c]	.07	.46
Not enjoy	47.15	.43	1.17	2.6[c]	1.03	.3	.28	.55
Drool	49.78	.41	1.42[a]	6.5[c]	1.53[a]	6.1[c]	-.36	.41
Dribbling nose	42.08	.51	1.34	4.0[c]	1.24	2.1[c]	-.15	.38
Excess saliva	53.18	.39	1.22	3.7[c]	1.31	4.2[c]	-.33	.48
Chew	48.57	.41	1.14	2.3[c]	1.08	1.0	-.19	.52
Clear throat	56.13	.40	1.04	.6	1.21	2.9[c]	-.35	.50
Stick throat	52.58	.40	1.07	1.3	1.20	2.8[c]	-.22	.52
Thick saliva	55.42	.40	1.12	2.0[c]	1.20	2.8[c]	-.31	.53
Cough stick throat	50.53	.40	1.02	.4	1.15	2.0[c]	-.22	.51
Gag	47.28	.43	1.00	.0	1.08	1.0	-.34	.48
Dribbling mouth	46.33	.44	1.06	1.0	1.07	.8	-.31	.48
Cough	54.93	.40	.89	-2.0[c]	1.04	.7	-.36	.53
Choke liquid	49.56	.41	.89	-2.0[c]	.90	-1.4	-.26	.55
Stick mouth	48.42	.42	1.05	.8	1.00	.1	-.24	.52
Choke food	49.11	.41	.82	-3.3[c]	.82	-2.6[c]	-.20	.58
Figure out	49.43	.41	.83	-3.1[c]	.83	-2.4[c]	.11	.60
Difficult find foods	48.02	.42	.95	-.9	.96	-.5	.13	.55
Hard time	48.89	.41	.95	-.8	.98	-.2	.09	.55
Speak clearly	51.84	.39	1.10	1.8	1.18	2.5[c]	.00	.53
Fear choke	46.56	.43	.89	-1.9	.82	-2.3[c]	.19	.59
Fear pneumonia	45.15	.45	1.27	3.7[c]	1.28	2.8[c]	.01	.44
Afraid choking	45.97	.44	.99	-.2	.87	-1.5	.12	.55
Never know choke	45.91	.44	1.18	2.6[c]	1.10	1.1	.07	.52
Depressed	47.67	.42	.98	-.3	.92	-1.0	.36	.59
Annoyed	52.72	.39	.80	-3.8[c]	.77	-3.6[c]	.54	.68
Discouraged	49.11	.41	.75	-4.7[c]	.69[b]	-4.5[c]	.58	.67
Frustrated	50.66	.40	.80	-3.9[c]	.77	-3.5[c]	.57	.66
Impatient	50.04	.40	.81	-3.7[c]	.80	-3.0[c]	.45	.64
Not go out	50.85	.40	.80	-3.8[c]	.74	-4.0[c]	.55	.67
Social life	50.67	.40	.69[b]	-6.1[c]	.66[b]	-5.4[c]	.62	.68
Change work	49.51	.40	.67[b]	-6.6[c]	.62[b]	-6.0[c]	.59	.69
Social gathering	50.32	.40	.66[b]	-6.8[c]	.62[b]	-6.2[c]	.61	.70
Role	48.28	.41	.78	-4.1[c]	.69[b]	-4.4[c]	.48	.63
Weak	51.48	.40	1.04	.7	1.14	2.0[c]	-.32	.50
Tired	54.47	.39	1.07	1.2	1.29	3.9[c]	-.48	.46
Exhausted	50.14	.40	1.08	1.4	1.29	3.8[c]	-.43	.44
Fall asleep	49.42	.40	1.42[a]	6.5[c]	1.61[a]	6.9[c]	-.41	.39
Stay asleep	51.91	.39	1.33	5.4[c]	1.64[a]	7.9[c]	-.41	.41

Notes. [a] Infit or Outfit MnSq >1.4; [b] Infit or Outfit MnSq <0.7; [c] Z-STD ≤-2.0 or ≥2.0

We next examined for misfit of items on each subscale (see Table 5). In examining for contradiction between infit and outfit scores (see Table 5), although there were no contradictions in the direction of scores, the Z-STD for two items on the Symptom subscale (Dribble nose, and Chew) and one item of the Mental health subscale (Annoyed) were underfitting on infit, but were not under fitting on outfit. Underfit (MnSq > 1.4; Z-STD > 2) was observed in one item of the Fear subscale (Fear pneumonia). The Mental health subscale had one item (Depression) with an infit MnSq that was close to underfit (1.39) and was under fitting (1.46) on outfit scores.

Both MnSq and Z-STD infit scores were outside the model (overfit- MnSq < 0.7; Z-STD < -2) for the Fear subscale (Afraid of choking). Overfit reported as Z-STD scores (< -2 for both infit and outfit) on items in the Symptom subscale (Choke food), the Mental Health subscale (Discouraged and Frustrated), the Social Functioning subscale (Change work and Social gathering), and the Fatigue subscale (Tired).

Misfit of items for the two factors were next examined (see Table 6). In examining for contradiction between infit and outfit scores, there are were no contradictions in the direction of scores. The underfit (Z-STD scores) on three Factor One items (Dealing, Distracting and Impatient) were considerably lower on infit compared to outfit.

Underfit of both infit and outfit scores (MnSq > 1.4; Z-STD > 2) was observed in Factor One (Longer, Don't care, No hunger and Fear pneumonia) and all but MnSq infit on two items (Speak clearly and Never know choke). Items that were more underfitting only on Z-STD scores (both infit and outfit) for Factor One was Forever and for Factor Two were Weak and Fall asleep.

Both MnSq and Z-STD infit and outfit scores were outside the model (overfit- MnSq < 0.7; Z-STD < -2) for the following items of Factor One: Social life, Change work, and Social gathering. Infit reported as Z-STD scores indicated overfit (< -2) on items in Factor One (Annoyed, Discouraged, Frustrated, Impatient, Not go out and Role) and on Factor Two (Tired and Exhausted).

Table 5 Individual item fit statistics and principal component analysis for subscales

				Infit		Outfit		Factor	Point biserial
Scales	**Items**	**Measure**	**SE**	**MnSq**	**Z-STD**	**MnSq**	**Z-STD**	**loading**	**Correlations**
Burden	Dealing	55.02	.93	.99	-.2	.97	-.3	.02	.94
	Distracting	44.98	.93	.99	-.1	.97	-.4	.02	.94
Eating duration	Longer	64.54	.86	1.01	.2	.97	-.3	.15	.92
	Forever	35.46	.81	.98	-.3	.95	-.5	.15	.91
Eating desire	Don't care	48.62	.58	1.15	1.8	1.03	.4	.94	.80
	No hunger	52.02	.56	.81	-2.7	.77	-2.9	-.14	.85
	Not enjoy	49.36	.58	1.08	1.0	1.00	.0	-.83	.80
Symptoms	Drool	49.38	.43	1.34	5.4[c]	1.37	4.8[c]	.63	.50
	Dribbling nose	40.69	.54	1.30	3.7[c]	1.10	.9	-.05	.46
	Excess saliva	53.30	.42	1.13	2.2[c]	1.13	2.0[c]	.72	.58
	Chew	48.07	.44	1.12	2.0[c]	1.05	.6	.06	.57
	Clear throat	56.71	.43	1.02	.3	1.10	1.5	-.08	.57
	Stick throat	52.54	.43	1.03	.6	1.04	.6	-.45	.59
	Thick saliva	55.92	.43	1.00	.1	1.02	.4	.48	.62
	Cough stick throat	50.23	.43	.96	-.7	.98	-.3	-.48	.59
	Gag	46.62	.45	.96	-.7	.96	-.5	-.10	.55
	Dribbling mouth	45.44	.46	.95	-.8	.88	-1.5	.35	.57
	Cough	55.43	.43	.85	-2.7[c]	.93	-1.2	-.30	.60
	Choke liquid	49.19	.43	.90	-1.9	.88	-1.8	-.35	.59
	Stick mouth	47.81	.44	.90	-1.7	.85	-2.1[c]	-.06	.61
	Choke food	48.68	.43	.80	-3.7[c]	.80	-3.0[c]	-.56	.62
Food Selection	Figure out	53.30	.88	.97	-.4	.83	-2.1[c]	.05	.92
	Difficult find foods	46.7	.90	1.00	.1	.89	-1.3	.05	.91
Communication	Hard time	44.12	.82	1.00	.1	.98	-.3	.06	.91
	Speak clearly	55.88	.80	.96	-.5	.92	-1.1	.06	.92
Fear	Fear choke	51.20	.60	.89	-1.5	.88	-1.4	-.52	.82
	Fear pneumonia	48.64	.61	1.57[a]	6.3[c]	1.59[a]	5.8[c]	1.00	.71
	Afraid choking	50.14	.60	.66[b]	-5.0[c]	.68	-4.3[c]	-.35	.84
	Never know choke	50.02	.60	.92	-1.1	.85	-1.8	-.48	.81
Mental health	Depressed	43.93	.67	1.39	4.7[c]	1.46[a]	4.6[c]	-.38	.81

				Infit		Outfit		Factor	Point biserial
Scales	**Items**	**Measure**	**SE**	**MnSq**	**Z-STD**	**MnSq**	**Z-STD**	**loading**	**Correlations**
	Annoyed	57.07	.65	1.15	2.0[c]	1.12	1.6	-.58	.86
	Discouraged	47.55	.66	.76	-3.5[c]	.71	-3.9[c]	-.19	.88
	Frustrated	51.55	.65	.73	-4.1[c]	.73	-3.8[c]	.58	.89
	Impatient	49.90	.65	.97	-.4	.97	-.3	.75	.87
Social functioning	Not go out	53.07	.71	1.25	3.2[c]	1.28	3.4[c]	.79	.86
	Social life	52.33	.71	.85	-2.1[c]	.88	-1.6[c]	.37	.89
	Change work	48.66	.72	.79	-2.9[c]	.78	-3.0[c]	-.66	.89
	Social gathering	51.23	.71	.71	-4.2[c]	.71	-4.1[c]	-.48	.90
	Role	44.71	.73	1.36	4.2[c]	1.31	3.7[c]	-.24	.84
Fatigue	Weak	48.24	.70	1.30	4.1[c]	1.27	3.9[c]	1.00	.83
	Tired	57.58	.69	.81	-3.0[c]	.81	-3.1[c]	-.61	.88
	Exhausted	44.18	.71	.90	-1.6	.85	-2.3[c]	-.61	.87
Sleep	Fall asleep	45.69	.76	1.00	.1	.95	-.7	.05	.90
	Stay asleep	54.31	.74	.97	-.4	.93	-.9	.05	.90

Notes. [a] Infit or Outfit MnSq >1.4; [b] Infit or Outfit MnSq <0.7; [c] Z-STD ≤-2.0 or ≥2.0

Table 6 Individual item fit statistics and principal component analysis for factors

				Infit		Outfit		Factor	Point biserial
Scales	**Items**	**Measure**	**SE**	**MnSq**	**Z-STD**	**MnSq**	**Z-STD**	**loading**	**Correlations**
Factor One	Dealing	55.23	.43	.72	-5.1[c]	.87	-1.7	0.07	.71
	Distracting	53.09	.43	.84	-2.8[c]	.87	-1.7	0.03	.68
	Longer	59.46	.45	1.48[a]	6.4[c]	1.80[a]	7.9[c]	0.49	.61
	Forever	51.50	.43	1.18	2.8[c]	1.16	1.9[c]	0.54	.60
	Don't care	46.84	.46	1.63[a]	8.0[c]	1.87[a]	7.5[c]	0.39	.46
	No hunger	48.83	.44	1.48[a]	6.6[c]	1.79[a]	7.3[c]	0.52	.50
	Not enjoy	47.17	.46	1.23	3.2[c]	1.15	1.6	0.41	.56
	Figure out	49.87	.44	.97	-.5	1.03	.4	0.36	.60
	Difficult find foods	48.24	.45	1.08	1.2	1.11	1.2	0.37	.56
	Hard time	49.17	.44	1.12	1.8	1.21	2.3[c]	0.01	.56
	Speak clearly	52.69	.43	1.33	5.0[c]	1.63[a]	6.7[c]	0.13	.55
	Fear choke	46.52	.46	.99	-.1	1.00	.1	-0.36	.57
	Fear pneumonia	44.93	.48	1.47[a]	5.9[c]	1.69[a]	5.6[c]	-0.08	.45
	Afraid choking	45.88	.47	1.10	1.5	1.08	.8	-0.32	.54
	Never know choke	45.80	.47	1.36	4.8[c]	1.63[a]	5.5[c]	-0.34	.50
	Depressed	47.80	.45	1.02	.3	1.06	.6	-0.51	.59
	Annoyed	53.74	.43	.76	-4.2[c]	.73	-3.7[c]	-0.32	.71
	Discouraged	49.44	.44	.71	-5.2[c]	.64[b]	-4.9[c]	-0.54	.67
	Frustrated	51.25	.43	.75	-4.5[c]	.71	-4.0[c]	-0.51	.68
	Impatient	50.52	.43	.83	-2.9[c]	.85	-1.9	-0.5	.65
	Not go out	51.52	.43	.76	-4.3[c]	.73	-3.7[c]	-0.02	.68
	Social life	51.28	.43	.62[b]	-7.3[c]	.63[b]	-5.3[c]	-0.16	.70
	Change work	49.92	.44	.62[b]	-7.2[c]	.58[b]	-5.9[c]	-0.13	.69
	Social gathering	50.86	.43	.62[b]	-7.3[c]	.60[b]	-5.6[c]	-0.18	.70
	Role	48.45	.45	.77	-3.9[c]	.70	-3.9[c]	-0.12	.64
Factor Two	Weak	50.04	.56	1.13	2.0[c]	1.19	2.7[c]	-.32	.75
	Tired	55.86	.55	.77	-3.8[c]	.79	-3.4[c]	-.63	.81
	Exhausted	47.36	.56	.82	-2.9[c]	.78	-3.4[c]	-.60	.80
	Fall asleep	45.94	.57	1.29	4.0[c]	1.19	2.6[c]	.71	.74
	Stay asleep	50.79	.55	1.03	.5	.98	-.2	.67	.78

Notes. [a] Infit MnSq or Outfit MnSq >1.4; [b] Infit MnSq or Outfit MnSq <0.7; [c] Z-STD ≤-2.0 or ≥2.0

Dimensionality

PCA of residuals was undertaken to examine the dimensionality of the overall scale with all 44 items combined in a single scale (see Supplementary Table 1). The Rasch dimension explained 40.5 % of the variance in the data and >40% is considered a strong measurement of dimension [23]. Of the 40.5 % explained variance the item measures (25.7%) explain more of the variance than the person measures (14.8%). The raw variance explained by the items was more than 3.5 times the variance explained by the first contrast (7.3%). The total raw unexplained variance (59.5%) has an eigenvalue of 44.0 and the eigenvalue of the first contrast (5.38) and the second contrast (3.57) indicate a second and third (and approximating a 4th) dimension based on eigenvalue of >3.

When the dimensionality of each subscale was examined, all the subscales with just two items (Burden, Duration, Food selection, Communication, and Sleep) reported no significant clustering on the first contrasts. While the remaining 3-5 item scales were reported on first and more contrasts there were no significant contrasts (all < 2 eigenvalues). While the first and second contrast of the Symptom subscale were <3 eigenvalue (cut off for second dimension), they were > 2, suggesting some clustering of items (> 2 eigenvalue) rather than just random noise.

In examining the dimensionality for the two factors, rather than being just one factor the contrast scores on the Factor One scale would indicate the presence of a second dimension (>3 eigenvalue) and two further clusters with the strength of > 2 eigenvalue. Factor Two reported no significant clustering on the first contrasts suggesting it is a unidimensional scale.

The person-item dimensionality map for all 44 items combined shows that: a) there were not enough easy and difficult items, b) that many people were not aligned against items, and c) that there were redundant items as evidenced with several items aligning at the same level of person ability (see Supplementary Figure 2). Items that favoured the top (more difficult items) were Thick saliva, Drool, Excessive saliva, Clearing throat, Dribbling mouth, Weak, Tired, Exhausted, Fall asleep and Stay asleep. Items that favoured the bottom (easiest items) were:

Distraction, Longer, Forever, Don't care, No hunger, Not enjoy, Change work, Social gatherings, Role, Depressed, Annoyed, Discouraged, Frustrated and Impatient.

For the subscales with only two items: Burden, Eating duration, Eating desire, Food selection, Communication, and Sleep there was very little separation in the difficulty of the two items except for the Eating duration subscale. For the subscales with more than two items there were still very little spread of items and some redundancies; a similar pattern was observed for all the 44 items combined scale with not having enough easy or difficult items.

The person-item dimensionality map for the two factors were very dissimilar. Factor One showed a better clustering of people, but marked item redundancy. Conversely Factor Two showed very few people against items.

Differential Item Functioning (DIF)

The DIF analysis enabled examination of potential contrasting item-by-item profiles associated with the following variables: a) language, b) having or not having a confirmed diagnosis of OD (OD vs. no OD), c) medical diagnoses (other than OD), and d) gender. The summary of the DIF analysis for all 44 items combined as one scale is presented in Supplementary Table 2 and revealed that only the item Dribbling mouth showed significant DIF for all variables. Significant DIF was observed on three variables (language, OD vs. no OD, and medical diagnosis) for the items Drool, Excessive saliva, Stick throat, and Impatient. Overall, for all items combined in a single scale, most items showed significant DIF for the variables language and medical diagnosis.

The summary of the DIF analysis for subscales is presented in Supplementary Table 3 and revealed that none of the items showed significant DIF for all variables. Similar to all items combined in a single scale, significant DIF was observed on three variables (language, OD vs. no OD, and medical diagnosis) for the items Drool, Excessive saliva, and Stick throat. However, the item Impatient now only showed significant DIF on OD vs. No OD. Overall, for the subscales, most items showed significant DIF for the variable language and to a lesser degree for medical diagnosis compared with all items combined in a single scale. More items (6 items) showed

significant DIF on the gender variable for the subscales compared with all items combined (1 item).

The summary of the DIF analysis for the two factors is presented in Supplementary Table 4 and, like the subscales, revealed that none of the items showed significant DIF for all variables. The item Impatient reappeared (similar to all items combined) as showing significant DIF on the three variables (language, OD vs. no OD, and medical diagnosis); Fear choke appeared under Factor One as another variable that showed significant DIF on the three variables.

The items Drool, Excessive saliva, and Stick throat that showed significant DIF on the three variables for both all items as a single scale and subscales were not included in the factors as the Symptom subscale items were not part of the factors. Overall and similar to all items combined in a single scale, most items showed significant DIF for the variables language and medical diagnosis. More items (4 items) showed significant DIF on the gender variable for the factors compared with all items combined (1 item).

Discussion

Rating scale

When examining the overall category disorder of the SWAL-QOL (5-point rating scale; 1-5), all items showed disordered thresholds for categories 4 to 5. Although the remaining categories were not disordered, potential gaps in the category descriptions were identified in categories 3-4. For the eleven subscales only limited category disorder was identified by misfit; however, all subscales showed at least one adjacent category gap in the category descriptions. When evaluating category disorder for both factors, none of the adjacent categories showed gaps in the category descriptions for Factor One (Dysphagia specific QOL), whereas at least one adjacent category showed gaps in the category descriptions for Factor Two (Generic QOL) indicating at some category disorder based on fit. Step disorder was identified for all items combined and the Symptom subscale. All of these findings indicate a need to add additional options to the response scale, however results from this study do not indicate an appropriate number of response options for increasing rating scale validly and further testing with larger scales is required. Alternatively, item descriptors could to be made more specific so that

patients can use the current 5-point scale with increased accuracy. In addition, questions need to be reformulated in such a way that all items use the same response option descriptions, as currently descriptors vary between scales.

Summary stats

The summary statistics for item and person ability for all 44 items combined was good; however, problems with internal consistency were identified based on misfitting MnSq outfit scores. For most subscales, summary fit statistics revealed low person reliability thus indicating that people were not separated into the minimum of two levels (i.e. high performers and low performers). Instruments should be able to differentiate respondents into more than one level, as it will inform clinical judgement as to which aspect of QOL are in need of greatest attention during management. This suggests that the small number of items contained within each subscale evaluate a too wide a spread of QOL levels, and are therefore not sensitive enough to differentiate whether respondents truly are "high" or "low" performers in that particular aspect of QOL. Therefore, the collapsing of subscales resulting in the remaining subscales having more items should be considered.

When retrieving data from the literature on internal consistency using the traditional Cronbach's alpha (CTT), high values for most subscales were found [29-31]. Although some disagreement exists about whether to use a recommended reliability standard of 0.80 [16] or 0.70 [29-31] for Cronbach's alpha, only one or two out of the total number of subscales failed to reach beyond the reliability standard for group-level research. These results seem far more positive than when compared with the outcome using Rasch analyses.

Item fit

Several items were misfitting when using the scale as a whole, or as factors or in subscales. For all items combined, the removal of underfitting and overfitting items is recommended as underfitting items distort the model and overfitting items suggest the measure to be functioning better than expected. Several items relating to the fear and social functioning subscale proved problematic regardless if investigated as a whole measure, by subscale of by factor. The number of misfitting items was lower when used in subscales compared to when used in the scale as a

whole or, surprisingly, in factors. Based on these item fit analyses, all items that are both misfitting and redundant should be removed.

Dimensionality

PCA of residuals to examine the dimensionality of the overall scale indicated that the SWAL-QOL includes in fact 4 or 5 factors (underlying concepts assessed by the measure), in contrast to the 3 factors as suggested by McHorney, Robbins [16] or the six factors as determined by Vanderwegen, Van Nuffelen [31]. The person-item dimensionality maps revealed several redundant items and the lack of sufficient easy and difficult items. Item fit needs to be reanalysed to support decisions on which items should be removed due to both misfit and redundancy. Due to the emergence of multiple factors confirmed in this analysis, and as the developers suggested, the SWAL-QOL should not be considered as a single scale that measures a single underlying concept; the use of an overall score for the whole scale is therefore not supported. As such, it is also not recommended to calculate a so-called total SWAL-QOL score for all 44 items, nor for the subscales or for the factors as described by the developers [16].

DIF

Differential Item Functioning analysis was used to examine the potential contrasting item-by-item profiles associated with language, presence or absence of confirmed diagnosis of OD, medical diagnosis and gender. Theoretically, DIF would be expected on some items for both variables referring to OD and medical diagnosis, but not to the variables language or gender. However, while cultural differences may explain DIF on some of the language items, the fact that most items showed significant DIF on language seems more likely an indication of problems with cross cultural validation. Previously, Bogaardt, Speyer [32] also referred to existing cultural differences when translating the SWAL-QOL from American English to Dutch. In addition, many items showed unexpected DIF on gender. Moreover, when using subscales, as intended by the developers, more DIF on both gender and language are observed compared to when using the scale as a whole or by factors.

IRT

This study used IRT to assess the psychometric properties of the SWAL-QOL to augment previous psychometric studies using CTT. Although the procedures and statistical results in CTT are relatively easy to interpret, we purposely chose to use IRT; CTT makes judgement on the performance of the test as a whole and the specific population included, whereas IRT assesses the reliability of each item and its contribution to the overall construct under review and is independent from the sample group [23]. Therefore, we used IRT as we considered the statistical arguments, which in fact were the main rationale for this study, to be superior to the advantages of easy interpretation. As a consequence, the results and conclusions of this study may be more challenging to comprehend and to link to clinical practice than when CTT is used.

Limitations

We did not compare results between CTT and IRT; future research may perform both CTT and IRT using the same set of data, thus explaining in to more detail the differences, concerns and advantages of applying different frameworks for developing measures and evaluating measurement properties. Future studies may also focus in more depth on contrasting item-by-item profiles following DIF analysis of variables such as language or gender. In addition, other psychometric properties, including responsiveness, should be addressed. Responsiveness examines the ability of an instrument to detect change over time in the construct being measured [33]. Responsiveness was outside of the scope of our study but is considered to be an essential psychometric property. Only measures that are responsive to change over time can provide valid and reliable outcome data in intervention studies. It is our opinion that the SWAL-QOL need to be redeveloped using the new insights gained from this study.

Conclusions and implications for practice

In general, studies using CTT described the SWAL-QOL as a reliable and valid self-report measure of HR-QOL in OD [16, 29-32, 34, 35]. Our findings using Rasch analyses indicate some serious problems in relation to the quality of its psychometric properties. Researchers tend to generalise CTT findings while presenting limited data on an instrument's reliability and validity. Moreover, researchers often do not consider the more recently developed IRT which is readily taken on in other areas of research, but still underutilised in many health disciplines. This is not

the first time that results of IRT were not in line with outcomes from CTT when considering the validity and the reliability of a patient self-report measure within the area of OD. Cordier, Joosten [36] described the psychometric characteristics of the Eating Assessment Tool (EAT-10), a commonly utilised screening measure for OD in at-risk populations as lacking. The authors recommended caution in the use of the measure and emphasised the need for more research evaluating the EAT-10 using Rasch analysis.

Different interpretations of how to apply the SWAL-QOL data can be found in the literature. Some authors determine a total SWAL-QOL score [29, 35], which is not supported by either the developers of the SWAL-QOL [16] or the results from our Rasch analyses. Therefore, implementing a cut-off score in clinical practice to identify patients with swallowing problems based on this total score [35] is not recommended. Further, it remains unclear how clinicians should interpret data on the individual SWAL-QOL subscales (or factors) in daily patient care. Confusion also exists about the actual total number of subscales of the SWAL-QOL; authors disagree on whether to consider the items on clinical symptoms an independent subscale, thus bringing the total number of subscales to eleven [31, 32]. Moreover, our findings revealed many items with low factor loadings indicating that these items do not support the overall construct to be measured, which potentially undermines the construct validity of the SWAL-QOL.

Based on the Rasch analysis of the SWAL-QOL data, we recommend future studies to elucidate the underlying constructs related to swallowing difficulties using IRT analyses, such as Rasch analysis, and for these constructs to be clinically meaningful. The SWAL-QOL will need to be redeveloped to meet international standards, such as those proposed by the COSMIN taxonomy of measurement properties and definitions for health-related patient-reported outcomes [37]. Use of the current version of the SWAL-QOL is not supported by findings from this study. Redevelopment of the SWAL-QOL should address the need for additional rating responses, increased specificity in item descriptors, a reformulation of item questions so that all items use uniform rating response descriptors, the removal of misfitting items and the generation of new easy and hard items for inclusion in the measure. Alternatively, new

instruments could be developed using IRT analyses and that meet international standards for instrument development.

References

1. Roden DF, Altman KW. Causes of dysphagia among different age groups: a systematic review of the literature. Otolaryngologic Clinics of North America. 2013;46(6):965-87.
2. Bhattacharyya N. The prevalence of dysphagia among adults in the United States. Otolaryngology--Head and Neck Surgery. 2014;151(5):765-9.
3. Takizawa C, Gemmell E, Kenworthy J, Speyer R. A Systematic Review of the Prevalence of Oropharyngeal Dysphagia in Stroke, Parkinson's Disease, Alzheimer's Disease, Head Injury, and Pneumonia. Dysphagia. 2016:1-8.
4. Alagiakrishnan K, Bhanji RA, Kurian M. Evaluation and management of oropharyngeal dysphagia in different types of dementia: a systematic review. Archives of gerontology and geriatrics. 2013;56(1):1-9.
5. Kertscher B, Speyer R, Fong E, Georgiou AM, Smith M. Prevalence of oropharyngeal dysphagia in the Nethelands: A telephone survey. Dysphagia. 2015;30(2):114-20.
6. Lancaster J. Dysphagia: its nature, assessment and management. British journal of community nursing. 2015.
7. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. Journal of Nursing Scholarship. 2005;37(4):336-42.
8. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: A conceptual model of patient outcomes. Journal of the American Medical Association. 1995;273(1):59-65.
9. Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes: John Wiley & Sons; 2013.
10. Roy N, Stemple J, Merrill RM, Thomas L. Dysphagia in the elderly: preliminary evidence of prevalence, risk factors, and socioemotional effects. Annals of Otology, Rhinology & Laryngology. 2007;116(11):858-65.
11. Garcia-Peris P, Parón L, Velasco C, De la Cuerda C, Camblor M, Bretón I, et al. Long-term prevalence of oropharyngeal dysphagia in head and neck cancer patients: impact on quality of life. Clinical Nutrition. 2007;26(6):710-7.
12. Plowman-Prine EK, Sapienza CM, Okun MS, Pollock SL, Jacobson C, Wu SS, et al. The relationship between quality of life and swallowing in Parkinson's disease. Movement Disorders. 2009;24(9):1352-8.
13. Keage M, Delatycki M, Corben L, Vogel A. A systematic review of self-reported swallowing assessments in progressive neurological disorders. Dysphagia. 2015;30(1):27-46.
14. 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.
15. Bergamaschi R, Crivelli P, Rezzani C, Patti F, Solaro C, Rossi P, et al. The DYMUS questionnaire for the assessment of dysphagia in multiple sclerosis. Journal of the neurological sciences. 2008;269(1):49-53.
16. McHorney CA, Robbins J, Lomax K, Rosenbek JC, Chignell K, Kramer AE, et al. The SWAL–QOL and SWAL–CARE outcomes tool for oropharyngeal dysphagia in adults: III. Documentation of reliability and validity. Dysphagia. 2002;17(2):97-114.
17. Silbergleit AK, Schultz L, Jacobson BH, Beardsley T, Johnson AF. The dysphagia handicap index: development and validation. Dysphagia. 2012;27(1):46-52.
18. Woisard V, Lepage B. The" Deglutition Handicap Index" a self-administrated dysphagia-specific quality of life questionnaire: temporal reliability. Revue de laryngologie-otologie-rhinologie. 2009;131(1):19-22.
19. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research. 2012;21:651-7. doi: 10.1007/s11136-011-9960-1.
20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60(1):34-42.

21. Fitzpatrick R, Davey C, Buston M, Jones D. Evaluation of patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998;2(14):1-74.
22. Hambleton KH, Jones RW. An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and their applications to test development. Educational Measurement: Issues and Practice. 1993;12(3):38-47.
23. Linacre JM. A user's guide to W i n s t e p s® Rasch-model computer programs: Program manual 3.92.0. Chicago, IL: Mesa-Press; 2016a.
24. Linacre JM. WINSTEPS Rasch measurement computer program: Version 3.92.0. Chicago, IL: Winsteps; 2016b.
25. Wright BD, Masters GN. Rating Scale Analysis. Rasch Measurement: ERIC; 1982.
26. Linacre JM. Investigating rating scale category utility. Journal of Outcome Measurement. 1999;3(2):103-22.
27. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurment in the human sciences. 3rd ed. New York, NY: Taylor & Francis; 2015.
28. Linacre JM. Detecting multidimensionality: which residual data-type works best? Journal of outcome measurement. 1998;2:266-83.
29. Finizia C, Rudberg I, Bergqvist H, Rydén A. A cross-sectional validation study of the Swedish version of SWAL-QOL. Dysphagia. 2012;27(3):325-35.
30. Lam PM, Lai CKY. The validation of the Chinese version of the Swallow Quality-of-Life Questionnaire (SWAL-QOL) using exploratory and confirmatory factor analysis. Dysphagia. 2011;26(2):117-24.
31. Vanderwegen J, Van Nuffelen G, De Bodt M. The validation and psychometric properties of the Dutch version of the Swallowing Quality-of-Life Questionnaire (DSWAL-QOL). Dysphagia. 2013;28(1):11-23.
32. Bogaardt H, Speyer R, Baijens L, Fokkens W. Cross-cultural adaptation and validation of the Dutch version of SWAL-QoL. Dysphagia. 2009;24(1):66-70.
33. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010;63(7):737-45.
34. Khaldoun E, Woisard V, Verin E. Validation in French of the SWAL-QOL scale in patients with oropharyngeal dysphagia. Gastroentérologie clinique et biologique. 2009;33(3):167-71.
35. Rinkel RN, Verdonck-de Leeuw IM, Langendijk JA, van Reij EJ, Aaronson NK, Leemans CR. The psychometric and clinical validity of the SWAL-QOL questionnaire in evaluating swallowing problems experienced by patients with oral and oropharyngeal cancer. Oral oncology. 2009;45(8):e67-e71.
36. Cordier R, Joosten A, Clavé P, Schindler A, Bülow M, Demir N, et al. Evaluation of the Eating Assessment Tool (EAT-10) psychometric properties using Rasch analysis. under review.
37. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology. 2010;10(1):22.