Differential Item Functioning of the Jaw Functional Limitation Scale

Swaha Pattanaik, DrPH, BDS
Mike T. John, DDS, PhD
Department of Diagnostic and Biological Sciences
School of Dentistry
University of Minnesota
Minneapolis, Minnesota, USA

Seungwon Chung, PhD
Department of Educational Psychology
College of Human Development
University of Minnesota
Minneapolis, Minnesota, USA

San Keller, PhD
American Institutes for Research, Washington, DC, USA.

Correspondence to:
Dr Swaha Pattanaik
Department of Diagnostic and Biological Sciences
University of Minnesota School of Dentistry
515 Delaware Street Southeast, Minneapolis, MN 55455-0348
Email: swahapattanaik@gmail.com

Submitted May 26, 2021; accepted January 11, 2022.

Aims: To assess the differential item functioning (DIF) of the Jaw Functional Limitation Scale (JFLS) due to gender, age, and language (English vs Spanish). **Methods:** JFLS data were collected from a consecutive sample of 2,115 adult dental patients from HealthPartners dental clinics in Minnesota. Participants with missing data were excluded, and analyses were performed using data from 1,678 participants. Whether the item response theory (IRT) model assumptions of essential unidimensionality and local independence held up for the JFLS was examined. Then, using Samejima's graded response model, the IRT log-likelihood ratio approach was used to detect DIF. The magnitude and impact of DIF based on Raju's noncompensatory DIF (NCDIF) cutoff value of 0.096, Cohen's effect sizes, and test (or scale) characteristic curves were also assessed. **Results:** Essential unidimensionality was confirmed, but locally dependent items were found on the JFLS. A few items were flagged with statistically significant DIF after adjustment for multiple comparisons. The NCDIF indices associated with all DIF items were < 0.096, and they had small effect sizes of ≤ 0.2. The differences between the expected scores shown in the test characteristic curves were little to none. **Conclusion:** The present results support the use of the JFLS summary score to obtain psychometrically robust score comparisons across English- and Spanish-speaking, male and female, and younger and older dental patients. Overall, the magnitude of DIF was relatively small, and the practical impact minimal. *J Oral Facial Pain Headache 2023;37:33–46. doi: 10.11607/ofph.3026*

Keywords: *differential item functioning, item response theory, jaw functional limitation scale, oral health, patient-reported outcome measures*

Restricted jaw mobility can impact essential day-to-day functions such as biting, chewing, swallowing, and speech. Certain oral conditions, trauma-related injuries, aging, and treatment such as major prosthetic rehabilitation can alter functional jaw movements.[1] Jaw functional limitation is one among many other patient-reported outcomes used to describe patient suffering.[2] It is commonly used in several disciplines, including dentistry. Severity of jaw functional limitation is typically measured by the Jaw Functional Limitation Scale (JFLS), a global measure of functional limitation.[1,3–5] Due to its global application, the JFLS scores of several language versions[1,3–4,6–8] need to be comparable to allow for meaningful comparisons.

Previous studies have shown that the JFLS is a reliable and valid measure.[3,4] However, measurement invariance (MI), an important statistical property, has not yet been investigated for the JFLS. MI:

> *"…requires not only that the measured constructs have same meaning across groups, but also that group comparisons of sample estimates (eg, means and variances) reflect true group differences and are not contaminated by group-specific attributes that are unrelated to the construct of interest."*[9]

The MI of scores is supported when differential item functioning (DIF) is not found.[10] DIF is detected when scale items behave differently across population subgroups (such as gender, age, language spoken,

race, ethnicity, religion, sexuality, education level, etc) after controlling for the attribute under investigation (eg, reported jaw function).[10] The Patient-Reported Outcomes Measurement Information System (PROMIS) initiative recommends DIF assessment in key demographic groups as a crucial step for developing and validating patient-reported health measures.[10] Doing so helps researchers know if observed group differences are based on biased estimates.

In the context of this paper, DIF analysis is necessary for valid patient subgroup comparisons of perceived jaw functional limitation. Hence, it was hypothesized that the JFLS item properties are equivalent with respect to age, gender, and language. DIF related to language was assessed because item meanings may change during translation or because the interpretation of item meaning may vary in different cultures. Ensuring equivalence of JFLS scores across cultural and demographic groups would enable health care providers and researchers to make valid group comparisons with respect to limitations in jaw function.[12] The aim of this study was therefore to investigate DIF due to gender, age, and language.

Materials and Methods

Study Population, Recruitment, and Data Collection

A total of 2,115 adult dental patients were recruited from HealthPartners dental clinics in Minnesota for the original cross-sectional study.[13,14] Secondary data from the original study were used, and 1,678 patients were included in the present analysis because they had sufficient information for characterizing the construct. Those with more than half of the items missing were dropped from the analysis. Those with missing data were similar in background characteristics compared to those without missing data (data not reported but available upon request). The not applicable (NA) response option was also considered as missing.

Recruitment for the original study occurred from July 2014 to April 2016. A consecutive sample of patients was recruited at each participating HealthPartners dental clinic. A battery of patient-reported outcome measures, including the JFLS, was mailed to the patients. They completed the questionnaires at home and sent them back to the HealthPartners Institute and were given $50 incentive for their participation. Both English and Spanish speakers were identified using the language indicator in the electronic dental and medical records. Spanish speakers could also be bilingual patients; these bilingual patients were identified by checking their country of origin in the electronic health record, and the patient was considered bilingual if their language was English or missing and if their country of origin was from the list of countries identified as Spanish-speaking: Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Spain, Uruguay, and Venezuela.[11] The bilingual patients were offered an English version instead of Spanish if they preferred; thus, in this manner, bilingual individuals self-selected English or Spanish.

This research was conducted in accordance with accepted ethical standards for human-patient research practice. It was reviewed and approved by the Institutional Review Board of the HealthPartners Institute in Minneapolis, Minnesota (registration A11-136). All of the participants completed an informed consent form before their enrollment.

Jaw Functional Limitation Scale

The details of the development of the JFLS have been published elsewhere[3,4] and are briefly summarized here. The JFLS was derived from the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) checklist[15] and the Mandibular Functional Impairment Questionnaire (MFIQ) and developed using Rasch methodology.[16] Ohrbach et al initially created a preliminary 8-item JFLS.[3] Functional limitation as a unidimensional construct emerged from both the factor analytic and Rasch measurement modeling results for this preliminary instrument.[3] Based on the preliminary 8-item instrument, the developers then created a 20-item JFLS and posited 3 separate constructs assessing limitations in mastication, jaw mobility, and verbal and emotional expression.[4] However, support for these subscales has been mixed,[4,6] and the way in which researchers derive JFLS subscale scores has also differed.[1,4,6] Thus, MI was evaluated for the total JFLS score[1,7,8] only (as the JFLS subscale scores are not as widely used as the JFLS total score). The item responses were on a 0- to 10-point scale, with 0 meaning no limitation and 10 meaning severe limitation. Since the study participants were from a general dental population, an NA response option was also provided to them if the difficulties with jaw function were not applicable to their oral condition.[1,4] In line with the approach used in the original study, this option was excluded from the present analysis. While the original scale is on an 11-point (0 to 10) scale, the item responses were recoded to a 5-point (0–4) scale for the present analysis. Often categories are collapsed into fewer options to deal with low frequencies in some response categories. Importantly, collapsing categories in the present study can be justified with the present statistical approach; ie, item response theory

(IRT) modeling using the graded response model.[17] Specifically, the following recoding scheme following the PROMIS guideline[10,18] was used: 0 = 0; 1, 2, or 3 = 1; 4,5, or 6 = 2; 7, 8, or 9 = 3; 10 = 4. This manner of collapsing categories allows for the avoidance of any imbalance in alternatives. This approach was also in line with how response options are grouped together for the pain rating scales.[19]

The English version of the JFLS was translated into Spanish following a rigorous 11-step PROMIS-recommended methodology for scale translation and harmonization across Spanish-language varieties. Simancas-Pallares et al described the translation process in more detail for the Orofacial Esthetic Scale (OES).[13] The OES was another questionnaire given to the same sample of patients as part of the same battery of patient-reported outcome measures.

Procedures and Statistical Approach

IRT-based methods were used to examine DIF. IRT is a psychometric theory that refers to a family of associated statistical models.[20] IRT models predict the probability of correct responses based on the respondents' position on the latent traits (eg, perceived jaw functional limitation) continuum and the properties (or parameters) of the items on the scale.[21] Before fitting the IRT models, it is important to check whether the data meets the model assumptions. These assumptions include: *(1)* unidimensionality, which means a set of items on a scale reflect a single latent trait or phenomenon; and *(2)* local independence, which means the relationship of the item responses to each other is accounted for by a single latent trait, and so the items are not statistically related to each other after the latent trait is accounted for or statistically held constant.[21] The two assumptions for the JFLS data were checked before performing the DIF analyses.

Model Assumptions and Fit

Unidimensionality.

To test whether the assumption of "essential" unidimensionality holds for the JFLS, exploratory factor analysis (EFA) was conducted using the iterated principal factor method.[22] Differences in the magnitude of the eigenvalues (or variance) between the factors were examined, and the proportion of variance was accounted for by the first factor.[23,24] Unidimensionality was considered if the first factors accounted for greater than 60% of the variance and if the ratio of the variance explained by the first and the second factor was above 4.[24]

Model fit was also assessed using the root mean square error of approximation (RMSEA)[25,26] and Tucker-Lewis Index (TLI)[27,28] based on the M_2 statistic, which is asymptotically chi-square distributed like a Pearson's X^2 statistic and the likelihood ratio statistic G^2 but has better calibration and power.[29] The following criteria were used to assess model fit: for RMSEA, < 0.05 indicates close fit; 0.05 to 0.08 indicates reasonable fit; > 0.08 to 0.10 indicates mediocre fit; and > 0.10 indicates poor fit and that the model is not recommended for use[25,26]; and for TLI values, > 0.95 indicates good model fit.[26]

Local independence.

Locally dependent items were identified using multiple criteria. The standardized correlations were examined among item score residuals once the variance due to the underlying latent trait was partialled out (Cramer's V; correlation residuals > 0.15) as reported in the multidimensional item response theory (mirt) package in R.[20,21,30] Items with inflated item slope parameter estimates with values > 6 were also identified.[31,32] It was further examined whether removing such items would make meaningful differences in the parameter estimates.[32]

Description of the model.

JFLS items have either a 0 to 10 numeric rating scale (original scale) or a 0 to 4 scale (used in analyses). Samejima's graded response model (GRM) was used for ordered polytomous (> 2) response categories to conduct the DIF analyses.[33,34] The GRM is an extension of the two-parameter logistic (2PL) model, which is often applied for dichotomous items.[35] The 2PL model is characterized by two item parameters or properties: discrimination (slope) and difficulty (or location). As the discrimination parameters of JFLS items differ, these differences were taken into account when modeling the data by including the second parameter (discrimination).[36] For instance, another IRT model, the Rasch model, only estimates the location parameter for each item and assumes that the discrimination parameter is constant across all items.[35] Hence, the GRM was more appropriate for the JFLS because the discrimination parameter estimates were varying across all its items.

DIF detection procedure: IRT log-likelihood ratio modeling.

IRT log-likelihood ratio (IRTLR) modeling was used for DIF detection based on a nested model comparison approach,[34,37] and the null hypothesis that JFLS item discrimination and difficulty parameters do not differ with respect to age, gender, and language was tested. Age was treated as a categorical variable and was divided into two groups: 22 to 55 years and 56 to 97 years (median: 56 years).

For each of the three analyses (age, gender, and language), the test for DIF was performed by constructing *(1)* a model wherein the parameters of all items were fully constrained between the comparison groups; and *(2)* a model wherein the parameters for each item in the item set were varied ("studied item") while the parameters for all other items were

constrained to be equal ("anchor items").[34,38] A modified "all-other" method was used where various iterative processes were involved. Since there was no prior information regarding DIF in the JFLS item set, an iterative process was used to select anchor items as an initial step. Specifically, DIF was tested for every item while treating the rest of the items as anchor items. Then, the final set of anchor items was identified using another iterative process called purification, where, with each iteration, an item displaying DIF is deleted and the model re-estimated absent that item. Such purification has been recommended because the presence of DIF leads to errors in estimation and ultimately errors in DIF detection.[34,39] Last, the studied items were retested relative to the final set of anchor items, which links the two groups in each DIF analysis.[34,38] For evaluation of the DIF significance testing, the Benjamini-Hochberg (B-H) procedure[40] was used to adjust the *P* values due to multiple comparisons. Note that a *P* value of .05 was used without any adjustment during selection of the anchor items to protect against type II error.

The item's discrimination, or slope parameter, which describes how well an item discriminates between individuals at different trait levels (ie, higher and lower levels of jaw functional limitation), was also examined.[35,41] A difference in this parameter indicates nonuniform DIF.[41] Nonuniform DIF implies that one group is more likely to endorse the item at certain trait levels, while at other trait levels, the other group is more likely to endorse the item.[35,41] Subsequently, the item's location or intercept parameters (d_1, d_2, d_3, and d_4) were examined. Differences in the location parameters but not the item's discrimination or slope parameter would indicate uniform DIF. Uniform DIF means that one group is consistently more likely than the other to endorse an item at each trait level or the DIF is in the same direction across the jaw functional limitation continuum.[35,41]

Evaluation of DIF magnitude and effect sizes.

The magnitude of DIF refers to the extent to which item performance differs between or among the groups conditional on the trait being examined.[42] The magnitude of DIF was assessed using the expected item scores. The expected item score for an item is computed as the sum of the probabilities of a response to each of the possible response options or categories. The average of the difference in the expected item scores can be quantified using the noncompensatory DIF (NCDIF) index.[43] This index is referred to as noncompensatory because it does not account for bias from other items in the scale.[43] The widely used cutoff value of 0.096 was adopted for the NCDIF.[36,44] As an effect size, the square root of NCDIF was used.[34] The NCDIF cutoff values were based on the most widely used criteria proposed by Raju et al.[43] Cohen's suggestions to evaluate the effect sizes were followed: 0.2 was considered a small effect size, 0.5 a medium effect size, and 0.8 a large effect size.[45]

Evaluation of DIF impact.

"Impact" refers to the influence of DIF on the scale score. This was assessed with expected total scores, which are obtained by summing up all items for a particular value of a latent trait by varying the item parameters for the DIF items between groups. The group differences were visualized via test (or scale) characteristic curves (TCCs), which provide the effect of DIF on the total score. Specifically, the impact was evaluated by plotting the TCCs for all items and for the items flagged as having DIF.[34,46]

Software.

All IRT analyses were performed using the mirt package in R.[30] For the unidimensionality assumption checking, the statistical software package Stata version 14.0 was used.[47] The DFIT (Raju's Differential Functioning of Items and Tests framework) package in R was used to calculate the NCDIF index.[48]

Results

Descriptive Analysis

Study participant characteristics.

Among the 1,678 participants used in the analysis, 1,443 were English speakers and 235 were Spanish speakers. There were 592 male and 851 female English-speaking participants, and 96 male and 139 female Spanish-speaking participants. The mean age of English speakers was 56.6 years, and of Spanish speakers was 42.6 years. A total of 18.7% of the participants (16.3% among English speakers and 33.6% among Spanish speakers) were in a government assistance program, indicating lower socioeconomic status. Table 1 provides descriptive statistics of the study participants and of the JFLS items. Table 2 provides descriptive statistics of the JFLS total scores for the total sample stratified by age, gender, and language.

Model Assumptions and Fit

Unidimensionality

EFA results showed that the first factor accounted for about 87% of the variability, and the ratio of the variance explained by the first to the second factor was greater than 6. The first eigenvalue was substantially larger than the rest of the values, indicating "essential" unidimensionality.[20] A value of M_2 (26) = 84.26 ($P < .001$) was obtained for a unidimensional model on the final 13 items. An RMSEA value of 0.042 with a 90% confidence interval (0.032, 0.052) and TLI value of 0.993 were also obtained. Both indices indicated the unidimensional model closely fit the examined data.

Table 1 Sociodemographic Characteristics of the Participants and Mean Scores for JFLS Items

	All participants (n = 1,678)	English speakers (n = 1,443)	Spanish speakers (n = 235)
Sociodemographic characteristics, n (%)			
Women	993 (59.2)	854 (59.2)	139 (59.2)
Men	685 (40.8)	589 (40.8)	96 (40.9)
Lower socioeconomic status*	313 (18.7)	234 (16.3)	79 (33.6)
Mean (SD) age, y	54.7 (16.1)	56.6 (15.9)	43.0 (12.4)
JFLS items, mean (SD) score			
JFLS 1: Chew tough food	2.4 (3.4)	2.4 (3.4)	2.4 (3.6)
JFLS 2: Chew hard bread	2.6 (3.5)	2.7 (3.5)	2.2 (3.4)
JFLS 3: Chew chicken	1.1 (2.3)	1.0 (2.2)	1.3 (2.5)
JFLS 4: Chew crackers	0.9 (2.0)	0.8 (1.9)	1.4 (2.6)
JFLS 5: Chew soft food	0.4 (1.3)	0.3 (1.2)	0.7 (1.8)
JFLS 6: Eat soft food requiring no chewing	0.3 (1.2)	0.2 (1.0)	0.5 (1.7)
JFLS 7: Bite from whole apple	1.5 (3.0)	1.5 (3.1)	1.2 (2.7)
JFLS 8: Bite into a sandwich	0.8 (2.1)	0.8 (2.1)	1.0 (2.4)
JFLS 9: Open wide enough to talk	0.4 (1.6)	0.4 (1.4)	0.9 (2.3)
JFLS 10: Open wide enough to drink	0.3 (1.3)	0.2 (1.1)	0.7 (2.0)
JFLS 11: Swallow	0.5 (1.6)	0.5 (1.4)	0.8 (2.1)
JFLS 12: Yawn	0.6 (1.7)	0.5 (1.6)	0.9 (2.4)
JFLS 13: Talk	0.5 (1.6)	0.4 (1.4)	0.9 (2.2)
JFLS 14: Sing	0.6 (1.7)	0.4 (1.5)	0.8 (2.2)
JFLS 15: Putting on a happy face	0.7 (1.9)	0.6 (1.7)	1.0 (2.4)
JFLS 16: Putting on an angry face	0.4 (1.5)	0.4 (1.4)	0.8 (2.2)
JFLS 17: Frown	0.3 (1.3)	0.3 (1.2)	0.7 (1.9)
JFLS 18: Kiss	0.5 (1.7)	0.4 (1.6)	0.9 (2.3)
JFLS 19: Smile	0.8 (2.1)	0.7 (2.0)	1.2 (2.7)
JFLS 20: Laugh	0.7 (2.0)	0.6 (1.8)	1.1 (2.7)

*Data pertaining to income, profession, etc, are difficult to obtain. Therefore, participation in a government assistance program data were used as an indicator of lower socioeconomic status.

Table 2 Descriptive Statistics for JFLS Summary Scores

JFLS sum score	Minimum value	25th percentile	Median	Mean	75th percentile	Maximum value
Women	0	0	3	17.5	22	200
Men	0	0	2	13.9	14	200
Younger (≤ 55 y)	0	0	2	17.7	20	200
Older (≥ 56 y)	0	0	2	14.3	20	183
English	0	0	3	15.1	20	200
Spanish	0	0	1	21.5	21	200

Local independence.

Results indicated violations of the local independence assumption, a prerequisite for applying the IRT-GRM model. The standardized residual correlations for some item pairs had values above 0.15 (Table 3). The item pairs of items 1 and 2, items 7 and 8, and items 19 and 20 showed excessive residual correlations, with values over 0.20. Additionally, the slope estimates for items 10 and 17 were exceedingly high, with values of 6.82 and 6.68, respectively. A substantial impact of local dependence was found when examined through IRT calibration. The local dependence was strong enough that it could have defined additional latent variable(s) and led to misinterpretations of the findings from the DIF analysis.[32] Hence, a sensitivity analysis was performed by removing the items showing local dependence sequentially, starting from the items with excessive residual correlations up to the items with less excessive residual correlations, leading to the exclusion of items 1, 3, 7, 10, 15, 17, and 19. Exclusion of these items did not affect the scoring, as a single overall JFLS score was considered rather than separate subscale scores.

DIF Analyses

Language.

Three items, "chew hard bread" (item 2), "open wide enough to bite into a sandwich" (item 8), and "open wide enough to talk" (item 9), were identified as the

Table 3 Residual Correlations of JFLS Items

	1	2	3	4	5	6	7	8	9	10
1:Chew tough food	–	0.42*	0.17*	0.13	−0.12	−0.14	0.13	−0.11	−0.15	−0.18*
2: Chew hard bread	–	–	0.15*	0.12	−0.10	−0.12	0.13	−0.11	−0.13	−0.14
3: Chew chicken	–	–	–	0.16*	0.11	−0.09	−0.08	−0.09	−0.10	−0.10
4: Chew crackers	–	–	–	–	0.12	0.10	−0.08	−0.09	−0.09	−0.09
5: Chew soft food	–	–	–	–	–	0.14	−0.11	−0.09	−0.09	−0.08
6: Eat soft food requiring no chewing	–	–	–	–	–	–	−0.13	0.10	−0.09	0.08
7: Bite from whole apple	–	–	–	–	–	–	–	0.22*	−0.12	−0.12
8: Bite into a sandwich	–	–	–	–	–	–	–	–	0.08	−0.09
9: Open wide enough to talk	–	–	–	–	–	–	–	–	–	0.11
10: Open wide enough to drink	–	–	–	–	–	–	–	–	–	–
11: Swallow	–	–	–	–	–	–	–	–	–	–
12: Yawn	–	–	–	–	–	–	–	–	–	–
13: Talk	–	–	–	–	–	–	–	–	–	–
14: Sing	–	–	–	–	–	–	–	–	–	–
15: Putting on a happy face	–	–	–	–	–	–	–	–	–	–
16: Putting on an angry face	–	–	–	–	–	–	–	–	–	–
17: Frown	–	–	–	–	–	–	–	–	–	–
18: Kiss	–	–	–	–	–	–	–	–	–	–
19: Smile	–	–	–	–	–	–	–	–	–	–
20: Laugh	–	–	–	–	–	–	–	–	–	–

	11	12	13	14	15	16	17	18	19	20
1: Chew tough food	−0.12	−0.14	−0.15	−0.15	−0.15	−0.14	−0.12	−0.13	−0.13	−0.11
2: Chew hard bread	−0.11	−0.13	−0.14	−0.14	−0.13	−0.14	−0.12	−0.13	−0.13	−0.11
3: Chew chicken	−0.09	−0.11	−0.10	−0.10	−0.08	−0.10	−0.11	−0.10	−0.08	−0.09
4: Chew crackers	−0.08	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	−0.09	−0.09	−0.10
5: Chew soft food	−0.08	−0.09	−0.10	−0.09	−0.09	−0.09	−0.09	−0.10	−0.09	−0.09
6: Eat soft food requiring no chewing	0.07	−0.08	−0.09	−0.09	−0.08	−0.08	−0.09	−0.10	−0.09	−0.10
7: Bite from whole apple	−0.09	0.14	−0.10	−0.10	−0.10	−0.11	−0.11	−0.09	−0.10	−0.11
8: Bite into a sandwich	−0.08	0.10	−0.09	−0.10	−0.09	−0.11	−0.12	−0.09	−0.09	−0.09
9: Open wide enough to talk	0.10	0.08	0.11	0.10	−0.10	−0.10	−0.10	−0.10	−0.09	−0.09
10: Open wide enough to drink	0.12	0.10	0.10	0.11	−0.11	0.13	−0.11	0.10	−0.11	−0.10
11: Swallow	–	0.12	0.10	0.10	−0.09	−0.09	0.11	−0.10	−0.10	−0.10
12: Yawn	–	–	0.10	0.11	−0.10	−0.10	−0.09	0.09	−0.10	−0.09
13: Talk	–	–	–	0.14	−0.09	−0.10	−0.10	0.10	−0.10	−0.10
14: Sing	–	–	–	–	0.10	0.13	−0.12	0.10	−0.10	−0.11
15: Putting on a happy face	–	–	–	–	–	0.14	0.13	0.11	0.19*	0.16*
16: Putting on an angry face	–	–	–	–	–	–	0.16*	0.13	−0.13	0.11
17: Frown	–	–	–	–	–	–	–	0.12	0.12	−0.11
18: Kiss	–	–	–	–	–	–	–	–	0.13	0.12
19: Smile	–	–	–	–	–	–	–	–	–	0.23*
20: Laugh	–	–	–	–	–	–	–	–	–	–

*Item pairs with standardized residual correlations above 0.15 violating local independence assumption.

studied items to be retested for DIF. Using the specified anchor items (items 4, 5, 6, 11, 12, 13, 14, 16, 18, and 20), only item 2 was identified as having significant uniform DIF after adjustment for multiple comparisons using B-H correction. The direction of DIF was such that English speakers who endorsed this item had worse jaw function than Spanish speakers who endorsed this item, which suggests that the item score did not provide the same measurement across languages. Table 4 shows the final parameter estimates (and their standard errors) and DIF tests for language. The NCDIF index of 0.072 was slightly lower than the cutoff (0.096). Based on Cohen's suggestions,[45] the effect size of 0.268 was considered small.

As shown in Fig 1, the expected total scale scores for English- and Spanish-speaking patients did not differ when the only item with DIF was included in the calculation of the total scores. Because only one item with language DIF was identified, a total score based on multiple DIF items was not able to be calculated, and so there is no plot showing total scores based on DIF items only.

Gender

Four items, including "open wide enough to bite into a sandwich" (item 8), "open wide enough to talk" (item 9), "yawn" (item 12), and "laugh" (item 20), were identified as the studied items to be retested for DIF due

Table 4 Comparison of Language Groups: Item Parameters and Standard Errors for the Anchor Items and Studied Items with DIF

Item name	Group	a	d_1	d_2	d_3	d_4	a DIF	d DIF
2: Chew hard bread*	English	2.09 (0.12)	0.36 (0.09)	−1.27 (0.10)	−2.11 (0.11)	−3.59 (0.15)	< 0.01	15.6
	Spanish	2.06 (0.25)	−0.26 (0.25)	−1.98 (0.30)	−3.08 (0.35)	−4.61 (0.45)	(1.000)	(0.006)
4: Chew crackers	English	2.80 (0.16)	−2.23 (0.15)	−4.20 (0.20)	−5.73 (0.25)	−7.88 (0.39)	NS, anchor item	
	Spanish							
5: Chew soft food (eg, macaroni, canned or soft fruits, cooked vegetables, fish)	English	3.32 (0.24)	−4.13 (0.27)	−6.68 (0.37)	−8.04 (0.45)	−9.72 (0.63)	NS, anchor item	
	Spanish							
6: Eat soft food requiring no chewing (eg, mashed potatoes, apple sauce, pudding, pureed food)	English	3.24 (0.27)	−5.13 (0.35)	−7.24 (0.46)	−8.56 (0.56)	−10.01 (0.76)	NS, anchor item	
	Spanish							
8: Open wide enough to bite into a sandwich	English	3.36 (0.21)	−2.81 (0.19)	−5.03 (0.26)	−6.14 (0.30)	−7.84 (0.38)	NS, no DIF	
	Spanish							
9: Open wide enough to talk	English	5.33 (0.44)	−6.65 (0.52)	−9.30 (0.68)	−11.10 (0.80)	−13.04 (0.94)	NS, no DIF	
	Spanish							
11: Swallow	English	2.76 (0.18)	−3.18 (0.19)	−5.24 (0.26)	−6.63 (0.34)	−8.13 (0.46)	NS, anchor item	
	Spanish							
12: Yawn	English	3.30 (0.22)	−3.64 (0.23)	−5.68 (0.30)	−7.02 (0.37)	−8.89 (0.49)	NS, anchor item	
	Spanish							
13: Talk	English	5.18 (0.41)	−6.02 (0.46)	−8.82 (0.61)	−10.86 (0.73)	−12.95 (0.89)	NS, anchor item	
	Spanish							
14: Sing	English	5.10 (0.42)	−6.07 (0.47)	−8.57 (0.61)	−10.51 (0.74)	−12.39 (0.87)	NS, anchor item	
	Spanish							
16: Putting on an angry face	English	4.03 (0.32)	−5.31 (0.38)	−7.45 (0.48)	−9.05 (0.58)	-10.51 (0.69)	NS, anchor item	
	Spanish							
18: Kiss	English	3.80 (0.28)	−4.71 (0.32)	−6.68 (0.40)	−8.05 (0.47)	−9.09 (0.54)	NS, anchor item	
	Spanish							
20: Laugh	English	3.06 (0.20)	−3.34 (0.21)	−4.90 (0.26)	−6.19 (0.31)	−7.20 (0.36)	NS, anchor item	
	Spanish							

*Significant DIF item ($P < .05$).

to gender. Specified anchor items (items 2, 4, 5, 6, 11, 13, 14, 16, and 18) were used after adjustment for multiple comparisons. Items 8 and 12 showed significant nonuniform DIF. Item 20 showed uniform DIF, meaning female individuals who endorsed this item had worse jaw function than male individuals who endorsed this item. Table 5 shows the final item parameters and DIF tests for gender. The NCDIF indices for all three items were < 0.001, clearly not reaching the cutoff (0.096). The corresponding effect sizes for items 8, 12, and 20 were .002, < .001, and < .001, respectively, representing small effect sizes.

As shown in Fig 2b, the impact of DIF due to gender on total scores based on just the four items exhibiting DIF was minimal. When the total scores were based on all items as shown in Fig 2a, no evident impact of DIF due to gender was observed.

Age

Two items, "chew hard bread" (item 2) and "swallow" (item 11), were identified as the studied items to be retested for DIF. Using the specified anchor items (items 4, 5, 6, 8, 9, 12, 13, 14, 16, 18, and 20),

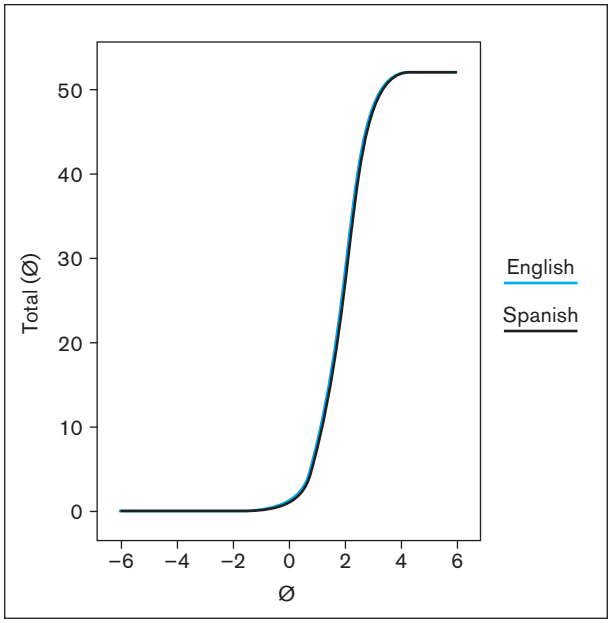

Fig 1 Expected total score of the JFLS by language group. There is only one plot because it was not possible to calculate a total score based on multiple DIF items, as only one item with language DIF was identified.

Table 5 Comparison of Gender Groups: Item Parameters and Standard Errors for the Anchor Items and Studied Items with DIF

Item name	Group	a	d_1	d_2	d_3	d_4	a DIF	d DIF
2: Chew hard bread	Female Male	2.02 (0.11)	0.53 (0.09)	−1.08 (0.10)	−1.94 (0.11)	−3.41 (0.14)	NS, anchor item	
4: Chew crackers	Female Male	2.83 (0.17)	−1.90 (0.15)	−3.85 (0.20)	−5.37 (0.25)	−7.49 (0.38)	NS, anchor item	
5: Chew soft food (eg, macaroni, canned or soft fruits, cooked vegetables, fish)	Female Male	3.38 (0.25)	−3.74 (0.25)	−6.29 (0.36)	−7.64 (0.44)	−9.31 (0.61)	NS, anchor item	
6: Eat soft food requiring no chewing (eg, mashed potatoes, apple sauce, pudding, pureed food)	Female Male	3.32 (0.28)	−4.77 (0.34)	−6.88 (0.44)	−8.19 (0.54)	−9.64 (0.74)	NS, anchor item	
8: Open wide enough to bite into a sandwich*	Female Male	3.16 (0.23) 4.16 (0.43)	−2.17 (0.19) −3.05 (0.35)	−4.21 (0.27) −5.77 (0.54)	−5.19 (0.31) −7.26 (0.64)	−6.87 (0.41) −9.06 (0.81)	6.3 (0.012)	12.5 (0.014)
9: Open wide enough to talk	Female Male	5.43 (0.46)	−6.05 (0.49)	−8.69 (0.64)	−10.48 (0.76)	−12.39 (0.89)	NS, no DIF	
11: Swallow	Female Male	2.83 (0.19)	−2.88 (0.19)	−4.93 (0.26)	−6.32 (0.33)	−7.82 (0.45)	NS, anchor item	
12: Yawn*	Female Male	3.12 (0.25) 4.11 (0.46)	−2.90 (0.23) −4.26 (0.47)	−4.86 (0.31) −6.58 (0.64)	−6.10 (0.39) −8.22 (0.79)	−7.99 (0.55) −10.07 (0.99)	5.2 (0.023)	11.2 (0.024)
13: Talk	Female Male	5.35 (0.43)	−5.50 (0.44)	−8.33 (0.59)	−10.37 (0.71)	−12.48 (0.87)	NS, anchor item	
14: Sing	Female Male	5.24 (0.43)	−5.54 (0.45)	−8.04 (0.59)	−9.98 (0.71)	−11.86 (0.84)	NS, anchor item	
16: Putting on an angry face	Female Male	4.13 (0.33)	−4.88 (0.36)	−7.01 (0.46)	−8.60 (0.56)	−10.06 (0.68)	NS, anchor item	
18: Kiss	Female Male	3.89 (0.29)	−4.29 (0.30)	−6.26 (0.39)	−7.62 (0.46)	−8.65 (0.52)	NS, anchor item	
20: Laugh*	Female Male	3.10 (0.24) 3.21 (0.33)	−2.87 (0.23) −3.23 (0.32)	−4.34 (0.29) −5.02 (0.43)	−5.83 (0.36) −5.88 (0.50)	−6.60 (0.41) −7.36 (0.63)	7.3 (0.660)	20.4 (0.046)

*Significant DIF item ($P < .05$).

items 2 and 11 showed significant uniform DIF after adjustment for multiple comparisons. Both items were a more severe indicator for the older participants. Table 6 shows the final item parameters and DIF tests for age. The NCDIF indices for items 2 and 11 were 0.010 and < 0.001, respectively, lower than the cutoff (0.096). The effect sizes for items 2 and 11 were 0.032 and 0.001, respectively, considered to be small effect sizes based on Cohen's suggestions.

Figure 3b shows total scores based on just the two items displaying DIF due to age. On this graph, the impact of DIF is noticeable. However, when total scores are based on the full set of items, as displayed by Fig 3a, DIF due to age did not have an evident impact.

Discussion

Patient-centered care is becoming increasingly prominent in health care research and practice. Patient-reported outcome measures play an important role in capturing patient experiences with oral health and overall health issues.[49] Discussing health outcomes with patients can help engage them in treatment decisions, which can improve their self-efficacy and relationship with their health care provider.[50] Patient-reported outcome measures are thus important for quality and performance measurement efforts to improve health care quality.[50]

DIF analysis helps assess measurement equivalence across diverse patient populations and thus paves a way for health care providers and researchers to better understand overall score differences.[51] Based on this hypothesis, the JFLS scores were equivalent for all studied groups. Few DIF items were detected that showed statistically significant DIF; however, the magnitude was small, and consequently these differences would have minimal and clinically irrelevant impact. Thus, researchers and clinicians can use the JFLS with confidence that the scores

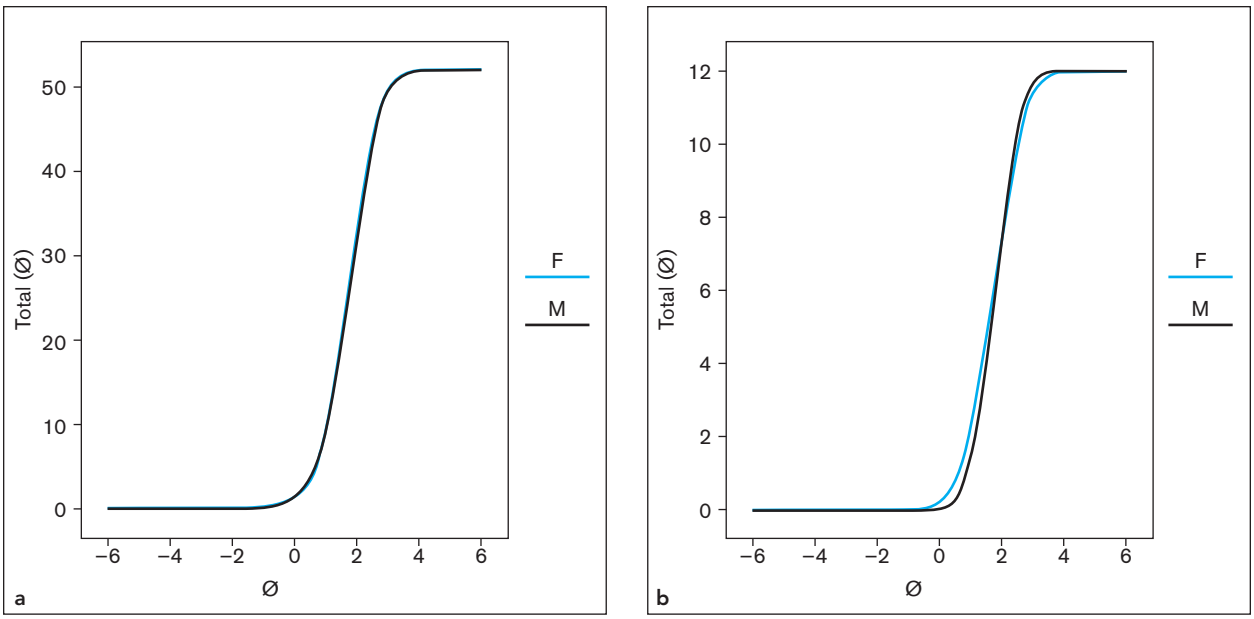

Fig 2 Expected JFLS total score by gender groups based on (a) all items and (b) the 4 items exhibiting DIF.

Table 6 Comparison of Age Groups (22–55 vs 56–97 y): Item Parameters and Standard Errors for the Anchor Items and Studied Items with DIF

Item	Group	a	d_1	d_2	d_3	d_4	a DIF	d DIF
2: Chew hard bread*	Older (≥ 56 y)	2.21 (0.16)	0.51 (0.12)	−1.19 (0.13)	−2.13 (0.15)	−3.43 (0.19)	6.6 (0.341)	20.4 (0.045)
	Younger (≤ 56 y)	1.86 (0.15)	0.21 (0.14)	−1.36 (0.15)	−2.15 (0.17)	−3.84 (0.22)		
4: Chew crackers	Old Young	2.77 (0.17)	−2.14 (0.16)	−4.11 (0.21)	−5.63 (0.26)	−7.76 (0.39)	NS, anchor item	
5: Chew soft food (eg, macaroni, canned or soft fruits, cooked vegetables, fish)	Old Young	3.29 (0.25)	−4.03 (0.27)	−6.57 (0.38)	−7.93 (0.46)	−9.61 (0.63)	NS, anchor item	
6: Eat soft food requiring no chewing (eg, mashed potatoes, apple sauce, pudding, pureed food)	Old Young	3.19 (0.28)	−5.01 (0.36)	−7.11 (0.46)	−8.41 (0.56)	−9.86 (0.76)	NS, anchor item	
8: Open wide enough to bite into a sandwich	Old Young	3.32 (0.22)	−2.70 (0.21)	−4.92 (0.27)	−6.02 (0.31)	−7.73 (0.39)	NS, anchor item	
9: Open wide enough to talk	Old Young	5.31 (0.46)	−6.52 (0.54)	−9.17 (0.69)	−10.97 (0.80)	−12.90 (0.94)	NS, anchor item	
11: Swallow*	Old	2.62 (0.24)	−2.71 (0.22)	−4.96 (0.34)	−6.66 (0.48)	−8.95 (0.92)	6.1 (0.234)	17.2 (0.002)
	Young	2.99 (0.29)	−3.70 (0.34)	−5.57 (0.43)	−6.78 (0.52)	−8.12 (0.64)		
12: Yawn	Old Young	3.30 (0.24)	−3.56 (0.24)	−5.62 (0.32)	−6.97 (0.38)	−8.84 (0.50)	NS, anchor item	
13: Talk	Old Young	5.15 (0.43)	−5.89 (0.47)	−8.70 (0.62)	−10.73 (0.74)	−12.84 (0.90)	NS, anchor item	
14: Sing	Old Young	5.12 (0.44)	−6.00 (0.50)	−8.50 (0.63)	−10.46 (0.76)	−12.36 (0.88)	NS, anchor item	
16: Putting on an angry face	Old Young	3.98 (0.33)	−5.19 (0.38)	−7.31 (0.49)	−8.90 (0.58)	−10.35 (0.69)	NS, anchor item	
18: Kiss	Old Young	3.81 (0.30)	−4.64 (0.33)	−6.62 (0.42)	−7.98 (0.49)	−9.02 (0.55)	NS, anchor item	
20: Laugh	Old Young	3.04 (0.21)	−3.25 (0.22)	−4.82 (0.27)	−6.10 (0.31)	−7.10 (0.37)	NS, anchor item	

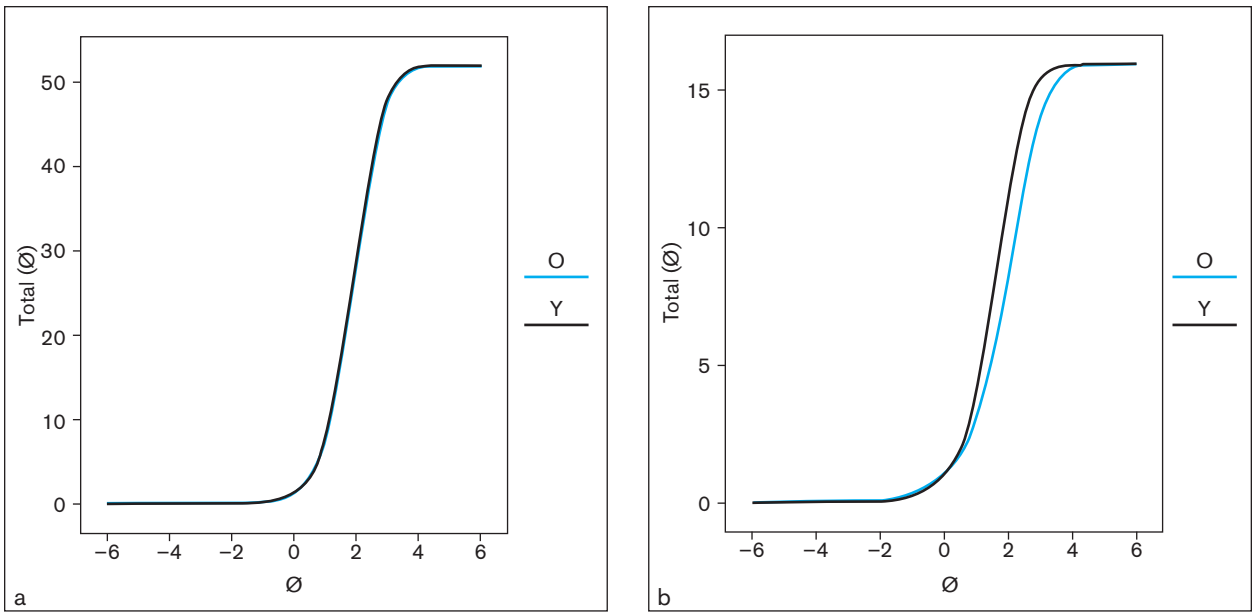

Fig 3 Expected JFLS total score by age group based on (a) all items and (b) the 2 items exhibiting DIF.

obtained are comparable, equivalent, and unbiased. The items reflect the construct of jaw functional limitation irrespective of gender, age, and language. Specifically, it was found that 5 out of 13 JFLS items exhibited DIF due to language, age, or gender; however, the overall magnitude and impact of including the DIF items in the total scores was small based on Raju's NCDIF cutoff values and Cohen's guidelines for evaluating effect sizes.[43,45] Minimal impact at the scale level was found when evaluating the effect of DIF on total scores by visually inspecting the TCCs for all items and for the items flagged as having DIF.

This is the first study to specifically assess DIF in the JFLS, and therefore the results are not able to be compared to the results of other studies. Few studies have performed DIF analysis of other patient-reported outcome measures in dentistry.[52–54] For example, in an earlier study, DIF across the English- and Spanish-language versions of the Orofacial Esthetic Scale (OES) were examined.[52] Similar to the present study, the IRTLR approach was used to detect DIF due to language for OES items. These findings evidenced measurement equivalence across the two language versions of the OES. Campos et al used structural equation modeling to investigate MI with respect to age, gender, socioeconomic status, and receipt of dental treatment across the Portuguese version of the OES (OES-Pt).[53] They found that the OES-Pt operated similarly to capture the concept of orofacial appearance in the comparison groups. In both studies, DIF assessment was performed on the full scale, as none of the OES items were locally dependent. Of note is that the OES is a relatively short questionnaire with 7 items assessing the appearance of orofacial structures. The present authors assume it presents fewer challenges in translation and analytical procedures compared to longer scales such as the JFLS. In another study, Lee et al assessed the equivalence of Oral Health Literacy Assessment (OHLA) scores between English- and Spanish-language groups.[54] They also used the IRTLR approach and found DIF was present in 23 out of 24 items, indicating a high level of item bias. The authors attributed DIF occurrences to cross-cultural differences and in how people from the two language groups perceive and understand oral health.[54] Language-related DIF was found with one item, "chew hard bread" (item 2), and a difference in expected total scale score between the two language groups was not observed when including this item. Again, the investigated outcomes differed: While the OHLA measures individuals' ability to obtain, process, and understand oral health information,[55] the JFLS targets the impact of oral diseases on functional limitation.

Strengths

The present study investigated regular dental patients—a highly relevant target population for the assessment of jaw functional limitation—using a large sample (n = 1,678). There are no established guidelines on the sample size requirement for DIF analysis,[56] as DIF detection methods vary. Typically, a sample size larger than 1,000 is deemed appropriate for DIF detection methods for polytomous

items.[57] IRT-based approaches were used for the present DIF assessment. IRT provides some advantages over confirmatory factor analysis (CFA), such as model flexibility, provision of useful psychometric statistics/applications, and modeling of more item parameters.[58] In terms of model flexibility, CFA/SEM assumes equal distance or an interval between response options (eg, the difference between 1 and 2 and 7 and 8 is assumed the same), whereas IRT does not necessarily make this assumption, allowing for greater model flexibility. Item and test information provide substantially more insight about psychometric properties, such as score accuracy, that vary across levels of the latent trait. In contrast, CFA/SEM assumes one constant value of score reliability for the entire scale. Finally, because IRT models each response category, more model parameters are estimated, allowing for a more detailed comparison between the investigated groups.

Limitations

The generalizability of these results will be affected by particular features of the present methods. The original study used a consecutive sampling approach, which may have introduced sampling bias, limiting the representativeness of the target population. The present authors did not specifically select a severely ill sample because the purpose was to evaluate the measurement properties of the JFLS in a general dental care population, which includes patients who are well and patients with a range of conditions. However, it is also to be noted that a key advantage of IRT over classical test theory (CTT) is its group-invariance principle.[59,60] In IRT, item parameters or properties are not dependent on the characteristics of the respondents; they are in fact invariant across samples given the IRT model fits the data. Bandalos[61] explained: ". . . if the IRT model fits, item parameter estimates are not dependent on the group from which they were obtained, and examinee ability estimates are not dependent on the particular items chosen to be on the test." Accordingly, the parameters of an item can be estimated from any group of subjects who have answered the item regardless of their position on the item response curve. For instance, even if moving from a group of participants with lower scores or less severe impairments (represented on the left side of the curve) to a population with higher scores or more severe impairments, such as TMD patients (represented on the right side of the curve), the item parameters should still be the same. Nonetheless, the present authors still recommend further empirical evidence regarding DIF in more severely ill patient populations.

Although several NCDIF cutoff values are available,[36,62,63] the present interpretations are based on the cutoff guideline used.[43] NCDIF interpretations can differ based on the researchers' use of a specific cutoff guideline. There are more effect size measures for DIF available (eg, signed/unsigned item difference in normal distribution, maximum difference in sample, expected score standardized difference) that offer detailed interpretations of DIF magnitude,[64] which needs more attention in future research evaluating DIF for scores from patient-reported outcome measures.

To examine age-related DIF, age was used as a categorical variable; particularly, groups of $\leq$ 55 years and > 55 years were used for analysis. Thus, conclusions could change if different age groups were used for analysis. To test this, a sensitivity analysis using a different categorization (> 65 and $\geq$ 65 years) was performed, and the results, while not identical (data not reported but available upon request), produced an effect size and impact on the scale scores that was small and similar to the results based on the age grouping in the present paper. Last, it is recognized that factors such as acculturation, education, income, and other indicators of socioeconomic status were not included in the DIF assessment and may have influenced the results. It is likely that many of the Spanish-speaking participants were bilingual and self-selected into the preferred language group. It is also recognized that acculturation may have had some influence on language-related DIF.[65] Although current study findings show that only one item displayed language-related DIF with small impact and magnitude, based on the available data, it is uncertain to what extent acculturation may have influenced the results. Level of education can also influence DIF and has not been investigated in the present study.[66] The present authors recommend future research studies to incorporate acculturation, education, and other indicators of socioeconomic status in their DIF assessment.

As for the scale itself, the dimensional structure of the JFLS and how responses of all items are scored have differed among prior studies.[1,3–8] While it is possible to score three separate subscale scores, a single overall score has been used most frequently in previous research and clinical application; thus, a single overall score was used.[1] Additional tests were performed to demonstrate an underlying unidimensional structure of the JFLS; this was important, as DIF may occur due to a misinterpretation of multidimensionality inherent in a measure. The present authors acknowledge that a reduced 13-item JFLS version was used for the DIF analysis; however, it was found that removing items with high residual correlations helped satisfy the assumption of local item independence. If this assumption is violated, any statistical analysis based on it would be misleading.[67]

Recommendations for Future Research

In terms of future research, presence of DIF is not a problem in itself, but is rather an opportunity to further study factors that led to its occurrence; for instance, examining how well the item showing language-related DIF has been translated and if it can be improved to better capture the intended meaning. The authors also recommend further scale refinement and evaluation efforts, including additional assessment of the JFLS dimensional structure and of excess item covariation. Further investigation of the JFLS dimensionality is important to the interpretation of DIF analysis results for that measure. Multidimensionality, and not bias in measurement, can be the cause of significant DIF results.[68] Researchers can apply the present methods to the analysis of JFLS data from other studies to determine whether any items violate assumptions of unidimensionality. Once found, those items can be dropped, and DIF analysis conducted on the remaining set. The DIF magnitude and impact were small, however, for caution, it is recommended that researchers carefully examine the individual DIF items for any possible response bias and evaluate whether the response bias results in clinically relevant problems. As different language versions of the JFLS are present, ideally, DIF assessment for other versions not examined here should be carried out, although it is presumed that other language versions developed with the same rigor as JFLS-English and -Spanish versions will likely not show substantial language-related DIF.

Evidence-based practice relies on patient-reported outcome measure scores in several disciplines of medicine[34,41,46] and dentistry.[69–71] These scores are also essential for value-based care.[72] The present results provide assurance to health care providers and researchers that the JFLS scores are equivalent across English- and Spanish-speaking, younger and older dental patients of male and female sex and allow for meaningful score comparisons across these patient subgroups.

Conclusions

While the JFLS needs more methodologic work, the present results suggest that its summary score allows psychometrically robust score comparisons across English- and Spanish-speaking, younger and older dental patients of male and female sex.

Highlights

- The JFLS has promising psychometric properties; however, DIF, while an important statistical property, has not yet been investigated for the JFLS. DIF occurs when items of an instrument show different measurement properties across groups of people who differ in background characteristics. The present study is the first to assess DIF due to gender, age, and language (English vs Spanish) in the JFLS using an IRT approach.
- It was found that JFLS scores are equivalent across English- and Spanish-speaking, younger and older dental patients of male and female sex and allow for meaningful score comparisons across these dental patient subgroups.

Acknowledgments

The National Institute of Dental and Craniofacial Research of the National Institutes of Health, USA, under Award Numbers R01DE022331 and R01DE028059, supported the study. The authors report no conflicts of interest.

Author contributions: S.P., M.J., S.C., and S.K.: study design, data analysis, interpretation, and drafting of the manuscript.

References

1. Oghli I, List T, John MT, Häggman-Henrikson B, Larsson P. Prevalence and normative values for jaw functional limitations in the general population in Sweden. Oral Dis 2019;25:580–587.
2. Mittal H, John MT, Sekulić S, Theis-Mahon N, Rener-Sitar K. Patient-reported outcome measures for adult dental patients: A systematic review. J Evid Based Dent Pract 2019;19:53–70.
3. Ohrbach R, Granger C, List T, Dworkin S. Preliminary development and validation of the jaw functional limitation scale. Community Dent Oral Epidemiol 2008;36:228–236.
4. Ohrbach R, Larsson P, List T. The jaw functional limitation scale: Development, reliability, and validity of 8-item and 20-item versions. J Orofac Pain 2008;22:219–230.
5. Schiffman E, Ohrbach R, Truelove E, et al. Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: Recommendations of the International RDC/TMD Consortium Network and Orofacial Pain Special Interest Group. J Oral Facial Pain Headache 2014;28:6–27.
6. Xu L, He Y, Fan S, Cai B, Fang Z, Dai K. Validation of a Chinese version of the Jaw Functional Limitation Scale in relation to the diagnostic subgroup of temporomandibular disorders. J Oral Rehabil 2020;47:1–8.
7. Lövgren A, Österlund C, Ilgunas A, Lampa E, Hellström F. A high prevalence of TMD is related to somatic awareness and pain intensity among healthy dental students. Acta Odontol Scand 2018;76:387–393.
8. Kapos FP, Look JO, Zhang L, Hodges JS, Schiffman EL. Predictors of long-term temporomandibular disorder pain intensity: An 8-year cohort study. J Oral Facial Pain Headache 2018;32:113–122.
9. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Med Care 2006;44(11 suppl 3):s78–s94.

10. PROMIS Cooperative Group. PROMIS® instrument development and validation scientific standards version 2.0. 2013:1-72. http://www.nihpromis.org/Documents/PROMISStandards_Vers2.0_Final.pdf?AspxAutoDetectCookieSupport=1.
11. Spanish Speaking Countries 2022. World Population Review. Accessed September 1, 2022. http://worldpopulationreview.com/countries/spanish-speaking-countries/
12. Harachi TW, Choi Y, Abbott RD, Catalano RF, Bliesner SL. Examining equivalence of concepts and measures in diverse samples. Prev Sci 2006;7:359–368.
13. Simancas-Pallares M, John MT, Prodduturu S, Rush WA, Enstad CJ, Lenton P. Development, validity, and reliability of the Orofacial Esthetic Scale - Spanish version. J Prosthodont Res 2018;62:456–461.
14. Simancas-Pallares M, John MT, Enstad C, Lenton P. The Spanish language 5-item oral health impact profile. Int Dent J 2020;70:127–135.
15. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord 1992;6:301–355.
16. Stegenga B, de Bont LG, de Leeuw R, Boering G. Assessment of mandibular function impairment associated with temporomandibular joint osteoarthrosis and internal derangement. J Orofac Pain 1993;7:183–195.
17. Andrich D. Distinctive and incompatible properties of two common classes of IRT models for graded responses. Appl Psychol Meas 1995;19:101–119.
18. Pattanaik S, John M, Chung S, Keller S. Comparison of two rating scales with the Orofacial Esthetic Scale and practical recommendations for its application. Health Qual Life Outcomes;2022:20:131.
19. Jacques E. (2020, January 5). 11 common types of pain scales. Verywell Health. Updated August 23, 2022. Accessed September 1, 2022. https://www.verywellhealth.com/pain-scales-assessment-tools-4020329
20. Lord FM, Novick MR, Birnbaum A. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.
21. Yang FM, Kao ST. Item response theory for measurement validity. Shanghai Arch Psychiatry 2014;26:171–177.
22. Browne MW. Fitting the factor analysis model. Psychometrika 1969;34:375–394.
23. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis, ed 8. Andover, UK: Cengage Learning, 2018.
24. Crins MHP, van der Wees PJ, Klausch T, van Dulmen SA, Roorda LD, Terwee CB. Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. PLoS One 2018;13:e0192187.
25. Maydeu-Olivares A, Joe H. Assessing approximate fit in categorical data analysis. Multivariate Behav Res 2014;49:305–328.
26. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS (eds). Testing Structural Equation Models. Newbury Park, CA: Sage, 1993:136–162.
27. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika 1973;38:1–10.
28. Cai L, Chung SW, Lee T. Incremental model fit assessment in the case of categorical data: Tucker-Lewis Index for item response theory modeling. Prev Sci 2021. Epub ahead of print May 10.
29. Joe H, Maydeu-Olivares A. A general family of limited information goodness-of-fit statistics for multinomial data. Psychometrika 2010;75:393–419.
30. Chalmers RP. Mirt: A multidimensional item response theory package for the R environment. J Stat Software 2012;48:1–29.
31. Chen WH, Thissen D. Local dependence indexes for item pairs using item response theory. J Educ Behav Stat 1997;22:265–289.
32. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16(suppl 1):5–18.
33. Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK (eds). Handbook of Modern Item Response Theory. New York, NY: Springer New York, 1997:85–100.
34. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychol Sci Q 2009;51:148–180.
35. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7:23–35.
36. Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous-DFIT framework. Appl Psychol Meas 1999;23:309–326.
37. Orlando Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. Application to the Mini-Mental State Examination. Med Care 2006;44(11 suppl 3):s134–s142.
38. Langer MM, Hill CD, Thissen D, Burwinkle TM, Varni JW, DeWalt DA. Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures. J Clin Epidemiol 2008;61:268–276.
39. Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. Med Care 2006;44(11 suppl 3):s152–s170.
40. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc: Series B (Methodological) 1995;57:289–300.
41. Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychol Sci Q 2008;50:538.
42. Teresi JA, Ocepek-Welikson K, Ramirez M, Fieo R, Fulmer T, Gurland BJ. Development of a short-form of the medication management test: Evaluation of dimensionality, reliability, information and measurement equivalence using latent variable models. J Nurs Meas 2018;26:483–501.
43. Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Appl Psychol Meas 1995;19:353–368.
44. Collins WC, Raju NS, Edwards JE. Assessing differential functioning in a satisfaction scale. J Appl Psychol 2000;85:451–461.
45. Cohen J. Statistical Power Analysis for the Behavioral Sciences, ed 2. Hillsdale, NJ: Erlbaum, 1988.
46. Hays RD, Calderón JL, Spritzer KL, Reise SP, Paz SH. Differential item functioning by language on the PROMIS physical functioning items for children and adolescents. Qual Life Res 2018;27:235–247.
47. StataCorp. Statistical Software: Release 14. College Station, TX: StataCorp LP, 2015.
48. Cervantes VH. DFIT: An R package for Raju's differential functioning of items and tests framework. J Stat Software 2017;76:1–24.
49. John MT. Health outcomes reported by dental patients. J Evid Based Dent Pract 2018;18:332–335.
50. Olde Rikkert MGM, van der Wees PJ, Schoon Y, Westert GP. Using patient reported outcomes measures to promote integrated care. Int J Integr Care 2018;18:8.
51. Walker CM. What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. J Psychol Assess 2011;29:364–376.

52. Pattanaik S, John MT, Chung S. Assessment of differential item functioning across English and Spanish versions of the Orofacial Esthetic Scale. J Oral Rehabil 2021;48:73–80.
53. Campos LA, Marôco J, John MT, Santos-Pinto A, Campos JADB. Development and psychometric properties of the Portuguese version of the Orofacial Esthetic Scale: OES-Pt. PeerJ 2020;8:e8814.
54. Lee J, Stucky B, Rozier G, Lee SY, Zeldin LP. Oral health literacy assessment: Development of an oral health literacy instrument for Spanish speakers. J Public Health Dent 2013;73:1–8.
55. Health Literacy in Dentistry. American Dental Association. Accessed September 1, 2022. https://www.ada.org/en/public-programs/health-literacy-in-dentistry
56. Scott NW, Fayers PM, Aaronson NK, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes 2010;8:81.
57. Kim S, Cohen A, Alagoz C, Kim S. DIF detection and effect size measures for polytomously scored items. J Educ Meas 2007;44:93–116.
58. Tay L, Meade AW, Cao M. An overview and practical guide to IRT measurement equivalence analysis. Organ Res Methods 2014;18:3–46.
59. Baker FB. The Basics of Item Response Theory, ed 2. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD: 2001.
60. Pattanaik S, John MT, Kohli N, et al. Item and scale properties of the Oral Health Literacy Adults Questionnaire assessed by item response theory. J Public Health Dent 2021;81:214–223.
61. Bandalos DL. Measurement Theory and Applications for the Social Sciences. New York, NY: Guilford Press, 2018.
62. Meade AW, Lautenschlager GJ, Johnson EC. A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. Appl Psychol Meas 2007;31:430–455.
63. Bolt DM. A Monte Carlo comparison of parametric and Nonparametric Polytomous DIF detection methods. Appl Meas Educ 2002;15:113–141.
64. Meade AW. A taxonomy of effect size measures for the differential functioning of items and scales. J Appl Psychol 2010;95:728–743.
65. Nguyen HT, Clark M, Ruiz RJ. Effects of acculturation on the reporting of depressive symptoms among Hispanic pregnant women. Nurs Res 2007;56:217–223.
66. Kim BS, Lee DW, Bae JN, et al. Effects of education on differential item functioning on the 15-Item modified Korean version of the Boston Naming Test. Psychiatry Investig 2017;14:126–135.
67. Baghaei P. Local dependency and Rasch measures. Accessed September 1, 2022.. https://www.rasch.org/rmt/rmt213b.htm
68. McDonald RP. A basis for multidimensional item response theory. Appl Psychol Meas 2000;24:99–114.
69. Reissmann DR. Dental patient-reported outcome measures are essential for evidence-based prosthetic dentistry. J Evid Based Dent Pract 2019;19:1–6.
70. Hua F. Increasing the value of orthodontic research through the use of dental patient-reported outcomes. J Evid Based Dent Pract 2019;19:99–105.
71. Palaiologou A, Kotsakis GA. Dentist-patient communication of treatment outcomes in periodontal practice: A need for dental patient-reported outcomes. J Evid Based Dent Pract 2020;20:101443.
72. Listl S. Value-based oral health care: Moving forward with dental patient-reported outcomes. J Evid Based Dent Pract 2019;19:255–259.