Improving the measurement of oral health-related quality of life: Rasch model of the oral health impact profile-14

Omara, M.; Salzberger, T.; Boecker, M.; Bekes, K.; Steiner, G.; Nell-Duxneuner, V.; ... ; Stamm, T.

Journal of Dentistry 114 (2021) 103819

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Improving the measurement of oral health-related quality of life: Rasch model of the oral health impact profile-14

Maisa Omara[a,b], Thomas Salzberger[c], Maren Boecker[d], Katrin Bekes[e], Guenter Steiner[b,f], Valerie Nell-Duxneuner[b,g], Valentin Ritschl[a,b], Erika Mosor[a], Margreet Kloppenburg[h], Judith Sautner[i], Bibiane Steinecker-Frohnwieser[j], Tanja Stamm[a,b,*]

[a] Section for Outcomes Research, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090, Vienna, Austria
[b] Ludwig Boltzmann Institute for Arthritis and Rehabilitation, 1090, Vienna, Austria
[c] Institute for Statistics and Mathematics, University of Economics and Business, Welthandelsplatz 1, 1020, Vienna, Austria
[d] RWTH Aachen University, Institute for Medical Psychology and Medical Sociology, Pauwelsstrasse 19, 52074, Aachen, Germany
[e] Department of Paediatric Dentistry, University Clinic of Dentistry, Medical University of Vienna, Sensengasse 2a, 1090, Vienna, Austria
[f] Division of Rheumatology, Department of Internal Medicine III, Medical University of Vienna, Vienna, Austria
[g] Klinikum Peterhof, Austrian Social Health Insurance Fund, Baden, Austria
[h] Department of Rheumatology, Leiden University Medical Center, Leiden, Netherlands
[i] LK Stockerau Korneuburg, 2. Med. Department, Lower Austria, Austria
[j] Ludwig Boltzmann Institute for Arthritis and Rehabilitation, Saalfelden, branch Gröbming, Austria

ARTICLE INFO

Keywords:
Rasch model
Measurement accuracy
Patient-reported outcome measures
Psychometric properties
Quality-of-life
Evidence-based dentistry

ABSTRACT

Objectives: The 14-item Oral Health Impact Profile (OHIP-14) is the most frequently used instrument to measure oral health-related quality-of-life (OHRQoL) in adults. Despite its popularity, its psychometric properties have been predominantly investigated based on the classical test theory while the fundamental principles of measurement have not been fully assessed. Therefore, our aim was to investigate to what extent the OHIP-14 meets the fundamental requirments of measurements.
Methods: We used the Rasch model to explore person-item-targeting, unidimensionality, local independence of items, invariance (differential-item-functioning, DIF), and the order of thresholds between response-options in the German version OHIP-14. We used data from osteoarthritis patients because hand disabilities and joint pain might influence oral hygiene. Furthermore, osteoarthritis in the temporomandibular-joint directly affects oral functioning.
Results: Five-hundred sixteen patients were included (mean age 66.5 years [±10.2; ranging from 34 to 89]; 71.3% [368] females). The OHIP-14 median total score was 0 (interquartile-range from 0 to 4), indicating a right-skewed distribution because many patients reported good OHRQoL. The instrument was found unidimensional. However, there was strong evidence of local dependency, disordered thresholds between response-options, and age-related DIF for item 5. A revised scoring scheme with three instead of five answer-options in all items and eliminating two items resolving local dependency, the newly adapted OHIP-12, showed better reliability and item-fit to the Rasch model than the original OHIP-14.
Conclusions: This study assesses, for the first time, the OHIP-14 in terms of fundamental principles of measurement and proposes an item-reduced OHIP-12 as a psychometrically more accurate version of the instrument.
Clinical significance: The Rasch model is essential to ensure instruments' precision and clinical meaningfulness when measuring OHRQoL in clinical practice and research. The OHIP-12, derived from the OHIP-14 by deleting two items due to local dependency, with a revised scoring scheme for all items distinguishing three answer-options instead of five, represents a psychometrically improved version of the instrument.

* Corresponding author at: Section for Outcomes Research, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria
E-mail address: tanja.stamm@meduniwien.ac.at (T. Stamm).

https://doi.org/10.1016/j.jdent.2021.103819
Received 20 May 2021; Received in revised form 23 August 2021; Accepted 13 September 2021
Available online 21 September 2021

1. Introduction

The importance of oral health-related quality-of-life (OHRQoL) has recently grown in clinical practice, oral health research, and dental community health programs. The need for a holistic approach to assess the impact of oral diseases by including physical, mental, and social well-being has sparked interest in measuring OHRQoL as one of the most important dental patient-reported outcomes [1]. The assessment of OHRQoL requires precise instruments which fully and accurately capture the patient perspective [1, 2].

While several instruments have been developed to measure OHRQoL, the Oral Health Impact Profile (OHIP) has become a widely used and cross-culturally accepted standard [3, 4]. Based on the WHO International Classification of Impairments, Disabilities, and Handicaps, Slade and Spencer (1994) developed a multidimensional framework and 49 items (the "original" OHIP-49) [3, 5]. Subsequently, shorter versions were proposed [3, 6, 7], with the OHIP-14 being the most frequently used version.

The OHIP-14 captures seven domains of OHRQoL with two items per domain: functional limitation, physical pain, psychological discomfort, physical disability, psychological disability, social disability, and handicap. All items are presented with a five-category rating scale ranging from "never" (0), "hardly ever" (1), "occasionally" (2), "often" (3), to "very often" (4). Even though the OHIP-14 was conceptualized as being composed of several domains, using one total score across all 14 items was recommended yielding a total score between 0 and 56 as the highest possible OHRQoL impairment. A series of empirical studies using OHIP-14 demonstrated that all items are indeed sufficiently related to justify one OHIP summary score characterizing OHRQoL [8, 9]. However, few studies investigating the structural dimensionality of OHIP using several methods in different populations, including healthy and patient populations found some evidence of multidimensionality [10, 11]. For example, Montero et al., who used exploratory and confirmatory factor analysis, reported three dimensions encompassing functional limitation, pain-discomfort, and psychosocial impacts as the underlying dimensions of OHRQoL [10].

To date, the properties of OHIP-14 have mainly been investigated and found satisfactory using classical test theory (CTT) [12] while modern test theory, specifically the Rasch model (RM) [13], with its more advanced potential for psychometric scale assessment, has not been applied yet. Besides the ability to test whether the items of a scale assess the same underlying construct (unidimensionality), the RM offers enhanced possibilities to investigate local dependency (LD) across items, measurement invariance, and the targeting of the scale in relation to the sample of interest.

Local dependency means that two items still have something in common even though the main underlying latent variable (OHRQoL) has been accounted for by the model. Thus, locally dependent items show correlated residuals (actual responses minus expected responses). Invariance means that item responses are related to the underlying latent variable in exactly the same way for different respondent groups. If invariance does not hold, differential item functioning exists, resulting in biased comparisons across groups. Moreover, in case the data meet the requirements of the RM, it is possible to infer linear measurement from an ordinal sum score allowing the use of parametric statistical methods [14-17].

Furthermore, the RM provides much more detailed insights into how items operate and whether person responses are meaningful [14]. It estimates item and person parameters independently: an item location parameter, also called item difficulty, expresses the item's endorsability (how much OHRQoL it takes to endorse the item), while a person location parameter reflects the respondent's level of OHRQoL [13]. Importantly, item difficulties should cover a broad range of OHRQoL matching the levels found in the population of interest, referred to as targeting. Good targeting increases the precision and interpretability of measures. In contrast, poor targeting means many extreme scores at the floor or the ceiling, implying a higher level of uncertainty and poor differentiation between respondents [18].

Forming a sum score across all items requires a unidimensional scale. Unidimensionality of an instrument consisting of multiple domains can still hold true, provided those domains are adequately related to one another [19].

Another fundamental requirement for self-reported measurement are locally independent items, i.e. the items scores have to be unrelated once the latent variable has been accounted for. Therefore, items sharing the exact same words or expressions, addressing the same facet of a construct, or logically building upon one another may violate local independence and should be considered for removal. Items with LD provide less information than locally independent items [20]. Moreover, LD across items might inflate reliability and introduce bias into parameter estimation.

Invariance is also fundamental requirement of measurement. It means that items operate the same way in different patients groups and their scores do not depend on personal factors [21]. For example, patients of the same level of OHRQoL but with more versus less osteoarthritis (OA) severity, or dentate versus edentulous respondents, may respond differently to an item. This phenomenon is called differential item functioning (DIF).

Furthermore, items have certain response options with so-called thresholds between them. A threshold indicates the location on the latent continuum where two adjacent response options (e.g. often and very often) are equally likely. The threshold estimates' order should be examined, and disordered thresholds suggest that some response options do not work as intended, e.g. the formulation is not well understood [22].

Since the properties mentioned above have not been investigated in the OHIP-14 so far. We tested the hypothesis that the OHIP-14 meets the fundamental requirements of measurements. Therefore, our study aimed at empirically investigating to what extent the OHIP-14 meets those requirements including, unidimensionality, locally independent items, DIF absence, and ordered threshold estimates as set out by the RM. Additionally, the person-item targeting was investigated in order to reveal how far the range of OHIP-14 item "difficulties" covers the level of OHRQoL of the study participants.

2. Materials and methods

2.1. Design

A cross-sectional psychometric study of the German version of OHIP-14 was conducted. We used data from patients with osteoarthritis to investigate how OA could affect the oral health of those patients. For instance, hand disabilities and joint pain might influence oral hygiene [23, 24] and OA in the temporomandibular joint directly affects oral functioning [25]. Additionally, researchers detected associations between radiographic knee OA and periodontitis [26]. Furthermore, OA is the most common musculoskeletal disease affecting an estimated 250 million people worldwide [27, 28]. The STROBE (Strengthening the Reporting of Observational Studies) for reporting cross-sectional studies were followed [29] (**Supplementary Table S1**).

2.2. Participants

Physician-diagnosed OA [30] patients aged ≥18 years with any form of OA in at least one joint, including hip, knee and hand joints, are enrolled in the so-called *Better Life with Osteoarthritis Registry (BLOAR)*, a multicentre OA registry in Austria. The responsible ethics committee approved the study (#2029/2016), and participants gave their oral and written informed consent. We used the first 632 consecutive patients as sample sizes around 500 were considered most appropriate for a reasonable interpretation of the RM's fit statistics [31]. Patients were recruited from six centers comprising rheumatology and orthopedic outpatient clinics as well as rehabilitation centers from three Austrian provinces (Vienna, Styria, and Lower Austria) with no predetermined dental conditions.

Only patients with complete data were included in the psychometric analysis to optimize the data for fit analysis, leaving a sample of 516 with no missing data for our analysis. Self-reported OA severity was investigated using a question ("*How bad would you say your arthritis is now?, with the potential answers of "mild," "moderate," or "severe"),* which was previously validated in osteoarthritis patients and found to be accurate for assessing patients' perceptions of their disease severity level [32].

2.3. Data analysis

Descriptive statistics were calculated, including measures of central tendency and variation. For the RM, the process of analysing data comprises the evaluation of model-data fit and testing how well the given observed data (item responses) meet the requirements of the RM, such as unidimensionality, the absence of LD and DIF. The Rasch model conceptualizes the data into a two-dimensional matrix, including item and person estimation. Hence, the RM predicts the probability of answering an item given a specific person estimate and tests whether the persons responding to the respective set of items (here to the OHIP-14) respond as expected under the model (test of fit). For example, a patient with OA who has a good OHRQoL would have a low probability of scoring high on an item reflecting high impairment of OHRQoL. In the following, the different indicators and test statistics used are described in more detail.

We examined the overall fit to the RM [33] and tested the fit residuals of each item. Non-significant fit residuals between −2.5 and +2.5 and non-significant item-trait interaction chi-square values were interpreted as item fit [15, 33].

The person separation index (PSI) was used as an indicator of reliability. The PSI is based on the same logic as Cronbach's alpha but uses linear measures and appropriate measurement error estimates that take the targeting into account. PSI's minimum acceptable value for investigations at the group level is 0.7, while for individual diagnostic, it is 0.85 [15, 34]. Moreover, we visually inspected the overall item-to-person targeting by using the item-person map on which item threshold estimates and person estimates are displayed on the same logit scale to examine how well the items matched the investigated sample.

To test unidimensionality, we used an approach proposed by Smith [19] namely the combination of principal component analysis (PCA) of the residuals followed by a series of t-tests to assess if subsets of items that loaded positively or negatively on the first principal component resulted in different estimates of person parameters suggesting multidimensionality. Unidimensionality is fulfilled if the lower 95% binominal confidence interval falls at 5% or below [19], which is the proportion of tests expected to be significant at that type-one error rate.

Local independence was assessed according to item residual correlations. A residual correlation of 0.2 above the average residual correlation was used as a cut-off to identify locally dependent items [15, 20]. In case LD was found, further investigations including subtest analysis and combining locally dependent items into one "super-item" were carried out to inform the possible deletion of redundant items in an iterative process.

We investigated DIF regarding several personal factors including gender (female/male), age-group (≤55, 56–65, 66–75 or ≥76 years), educational status [35], dentition (dentate/edentulous) and self-reported OA severity (mild, moderate, severe). If DIF existed in an item for a person factor with more than two categories (e.g. age-groups and OA severity), we determined the nature of those differences using posthoc analysis of the residual means [36].

To detect potentially disordered thresholds, category probability curves for each item (a graphic representation of how patients responded to the item answer options) were assessed. In the case of disordered thresholds, we first rescored the item by collapsing the answer options depending on the visual inspection of the category probability curves and each item's answer options' clinical meaning. A transformation of the instrument́s raw scores to linear equal-interval measures was also performed.

We used IBM SPSS Statistics 24, RUMM 2030 and the eRm package in R (www.r-project.org) [37].

3. Results

Five-hundred sixteen patients were included (mean age 66.5 years [± 10.2; ranging from 34 to 89]; 71.3% [368] females). Participant characteristics including their oral health profile, are summarized in **Table 1**. The participants reported slight OHRQoL impairments as the OHIP-14 median total score was 0 (interquartile range from 0 to 4), indicating a right-skewed distribution. More than half of the participants, 54,5% (281), had a total score of 0 (floor effect of very good OHRQoL). Moreover, the answer options distribution for each item of OHIP-14 was also depicted (**Supplementary Table S2**).

3.1. Consecutive analyses of fit to the rasch model

We started with an initial dataset which was then subsequently adapted and re-assessed according to the findings. Model fit statistics are shown in **Table 2**. The *initial dataset* showed misfit to the RM as indicated by a significant chi-square probability of item-trait interaction ($p = 0.00004$). While Cronbach's alpha was high (0.9), the PSI was only 0.3 due to the scale's poor targeting (**Table 2**). Items 6 (*Had to interrupt your meals)*, 2 *(Taste is impaired),* 1 *(Difficulties pronouncing certain words)* and 10 *(Completely incapable of doing anything)* were the most unlikely items to be affirmed by the respondents (**Table 3**).

Table 1
Descriptive characteristics of the OA sample, including oral health characteristics.

Characteristic	n (%)
Total number of participants: 516	
Gender - female	368 (71.3)
Age group (years)	
≤55	83 (16.1)
56–65	139 (26.9)
66–75	187 (36.2)
≥76	107 (20.7)
Educational level	
ISCED 1–2	396 (76.7)
ISCED 3	70 (13.6)
ISCED 4	14 (2.7)
ISCED 5	30 (5.8)
ISCED 6	6 (1.2)
Self-reported osteoarthritis severity	
Mild	99 (19.2)
Moderate	242 (46.9)
Severe	175 (33.9)
Smokers	257 (49.8)
Edentulous patients	116 (22.5)

Note: OA = osteoarthritis.
The international standard classification of education (ISCED) [35].
ISCED Level (1–2): Compulsory school, (Lower) Secondary education first stage.
ISCED Level 3: (Upper) Secondary education second stage (High school) Matura.
ISCED Level 4: Other post-secondary (not tertiary) education.
ISCED Level 5: First level tertiary education: University, first completed study course, Diploma studies, Bachelor and/ or Master level, MD, University of Applied Sciences.
ISCED Level 6: Second level tertiary education: PhD level, research qualification (University, doctoral studies based on a separate previous completed study course).

Table 2
Overall summary fit statistics to the Rasch model for the initial, rescored, and final revised datasets.

Dataset type *(N = 516)*	Mean item location (± SD)	Mean item fit residual (± SD)	Mean person location (± SD)	Mean person fit residual (± SD)	Chi-square probability of item-trait interaction	Person separation index (PSI)		Cronbach's α		Unidimensionality		
						With extremes	Excluding extremes	With extremes	Excluding extremes	N Significant tests	Percentage of significant t-tests	Lower bound of the 95% confidence interval
Initial dataset (OHIP-14)	0.0 (± 0.30)	−0.47 (± 1.49)	−2.51 (±1.09)	−0.34 (± 1.08)	**0.00004***	0.32	0.66	0.91	0.86	12	5.1%	2.3%
Rescored dataset (OHIP-14) *(rescored into three response categories)*	0.0 (± 0.48)	−0.59 (± 1.58)	−3.13 (±1.38)	−0.35 (± 1.09)	**0.002***	0.43	0.71	0.91	0.86	15	6.4%	3.6%
Final revised dataset (OHIP-12) *(rescored into three response categories and deleting items 6 and 10)*	0.0 (± 0.44)	−0.46 (± 1.54)	−2.96 (±1.35)	−0.31 (± 1.10)	0.006	0.40	0.67	0.90	0.83	11	4.7%	1.9%

Extreme values refer to patients with total score of OHIP-14 = 0 (N = 281, 54,5%).
The scale is considered unidimensional if the lower bound of the 95% confidence interval of the percentage of significant t-tests is ≤ 5%. [19].
* Bonferroni adjusted significance level at 0.004. The significant Chi-square probability of item-trait interaction indicates a misfit to the Rasch model. There is a deviation between the observed data and what is expected from the model.

We detected LD in eleven items within the *initial dataset*, and thirteen (92.9%) out of fourteen items had disordered thresholds. In principle, LD among items can be accounted for by forming subtests, which combine locally dependent items into one "super-item". However, subtest analysis did not provide a remedy in the present case suggesting that the reasons for LD were more complex. After that, we inspected threshold ordering for each item. A rescoring of the five response options by collapsing some categories resulting in a scoring scheme of 0 to 2 (three response levels) instead of 0 to 4 ensured a properly ordered threshold structure for all items. The rescoring's meaningfulness was supported by an improved total fit (chi-square item-trait interaction p = 0.002) in the new *rescored dataset*. The PSI also showed a higher value than the *initial dataset* (0.4) (**Table 2**).

However, rescoring did not eliminate the LD. Therefore, we scrutinized the clinical meaningfulness, the substantive content of the items displaying LD, their statistical properties, including item fit, the level of likelihood to endorse the item, and item residual correlations values (**Table 3**). Accordingly, we decided to delete item 6 (*Had to interrupt your meals)* because of LD with item 7 *(Uncomfortable eating certain foods),* item 5 *(Have a tense feeling*), and item 14 *(Feeling of uncertainty)*. Furthermore, we decided to delete item 10 (*Completely incapable of doing anything*) because of LD with item 9 *(Difficult to pursue your daily activities)* and item 11 *(Feel a little embarrassed)*. Importantly, item deletion did not result in a lack of content coverage as the omitted items were considered conceptually redundant. After the item deletion, good model fit was achieved (non-significant chi-square item-trait interaction p = 0.006 based on a Bonferroni-adjusted type-one error rate) (**Table 2**), and the PSI was not impacted (0.4). These findings supported the item deletion.

To sum up, our *final revised dataset* represents the proposed OHIP-12 with overall fit to the RM, three collapsed answer options for all items and a lower 95% confidence interval of 1.9% of significant t-tests (**Table 2**), indicating that OHIP-12 is unidimensional. However, targeting was still suboptimal as the range of item threshold parameters did not match the observed range of person parameters very well (**Fig. 1**). Some indication of possible LD remained for six items with residual correlations just slightly above the relatively strict limit we imposed (**Table 3**).

Regarding the individual item fit, only one item (item 5 *"Have a tense feeling"*) showed misfit to the RM and fell outside the pre-specified range of ± 2.5 with a fit residual value of −3.1. Moreover, the same item showed DIF for the person factor age group (**Table 3**). A posthoc analysis was performed and revealed that patients with ≥76 years of age answered *item 5* systematically differently than patients in age groups from 56 to 65 and 66–75 years. The impact of this DIF was evaluated by computing equated scores [36]. As the total score difference between age-groups ≥76 and <76 never exceeded one score point across the entire OHRQoL scale, this DIF was judged not meaningful **(Supplementary Figure S1 & S2, Table S3).** Regarding the other investigated personal factors, no DIF was detected, including the personal factor differentiating between the dentate and edentulous patients.

We performed the OHIP-12 raw score transformation to a linear equal-interval measure for future research to enable homogeneous estimates for various comparisons and calculations (**Supplementary Table S4**).

4. Discussion

This study investigated the psychometric properties of the OHIP-14 using the RM in a multicentre OA registry. It adds to the existing evidence and builds the basis for more accurate, consistent, and clinically meaningful dental patient-reported outcome measurement. While the OHIP-14 is the most widely used instrument worldwide to measure OHRQoL [3, 4], this study is the first to investigate extensively the psychometric properties of OHIP-14 based on an item- and scale-level in patients with the most common joint disorder, OA [28]. Our analysis proposes a solution that accommodates the present sample of respondents.

The psychometric properties of OHIP-14 have mainly been investigated and found satisfactory using classical test theory. [12] However, the RM provides more detailed insights on an item-based level. [14] Therefore, through using the RM some psychometric shortcomings could be revealed and some of them could be accounted.

The main psychometric issues of OHIP-14 in our study were the number of locally dependent items (11 items); the insufficient person-item targeting for the studied patient group; the misfit of item 5 as the

Table 3
The individual items characteristics of the *initial dataset* (D1) and the *final revised dataset* (D2) ordered by their location in the logit scale in D1.

Item number	Item description	Item Domain	Location		Standard error		Fit residual		ChiSq		ChiSq p-value		F-stat		F-stat p-value		Local Dependency		DIF	Final Recommendation (D2)
			D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1,D2	
4	Difficulties to relax	Psychological Disability	−0.563	−0.855	0.071	0.114	0.884	0.353	5.115	9.31	0.276	0.054	1.171	2.435	0.324	0.048	Item 3	Item 3		Rescored into three response categories instead of five, from 0,1,2,3,4 to 0,1,2
13	Pain in the mouth	Physical Pain	−0.407	−0.593	0.074	0.121	2.023	2.017	15.817	9.758	0.003	0.045	3.302	2.248	0.012	0.065	–	–		
14	Feeling of uncertainty	Psychological Discomfort	−0.273	−0.335	0.078	0.124	−1.508	−1.825	10.626	9.08	0.031	0.059	2.732	2.635	0.03	0.035	Item 6	–		
7	Uncomfortable eating certain foods	Physical Pain	−0.232	−0.225	0.076	0.124	1.742	1.739	4.141	5.531	0.387	0.237	0.754	1.156	0.556	0.331	Item 6	–		
3	Life was generally less satisfying	Handicap	−0.198	−0.064	0.082	0.13	−1.53	−1.644	8.759	5.787	0.067	0.216	2.006	1.553	0.094	0.188	Item 4	Item 4		
5	Have a tense feeling	Psychological Discomfort	−0.135	−0.135	0.08	0.127	**−2.624***	**−3.098***	12.819	13.963	0.012	0.007	4.089	5.353	0.003	0	Item 6		Age-group	
9	Difficult to pursue your daily activities	Social Disability	−0.012	0.357	0.086	0.141	−1.138	−1.377	4.217	3.713	0.377	0.446	0.805	0.839	0.523	0.502	Item 10 Item 8 Item 12	Item 8		
8	Irritable towards other people	Social Disability	0.072	0.359	0.087	0.14	−0.619	−0.905	5.245	4.045	0.263	0.4	1.115	0.942	0.35	0.44	Item 9	Item 9 Item 12		
12	Nutrition has been unsatisfactory	Physical Disability	0.108	0.316	0.089	0.14	−1.042	−0.935	3.208	2.088	0.524	0.72	0.581	0.377	0.677	0.825	Item 9	Item 8 Item 11		
11	Feel a little embarrassed	Psychological Disability	0.173	0.229	0.09	0.136	−1.253	−0.792	12.201	3.909	0.016	0.418	3.24	0.814	0.013	0.518	Item 10	Item 12		
10	Completely incapable of doing anything	Handicap	0.182		0.094		−0.731		9.022		0.061		2.367		0.054		Item 9 Item 11	–		Deleted
1	Difficulties pronouncing certain words	Functional Limitation	0.22	0.46	0.092	0.143	1.233	1.172	6.726	7.623	0.151	0.106	1.253	1.707	0.289	0.149	–			Rescored into three response categories instead of five, from 0,1,2,3,4 to 0,1,2
2	Taste is impaired	Functional Limitation	0.343	0.488	0.093	0.145	0.432	−0.206	1.464	1.267	0.833	0.867	0.34	0.309	0.851	0.872	–			
6	Had to interrupt your meals	Physical Disability	0.722		0.102		−2.439		8.385		0.078		3.61		0.007		Item 14 Item 7 Item 5	–		Deleted

D1: "*initial dataset*", **D2:** "*final revised dataset*" after adaptation, **DIF:** differential item functioning.
The rescoring of answer options into three response categories instead of five, from 0,1,2,3,4 to 0,1,2 was carried out as follows: the five-categories ranging from "never" (0), "hardly ever" (1), "occasionally" (2), "often" (3), to "very often" (4), were collapsed as "never" (0), "hardly ever" (1) and "occasionally" (1) together, as well as "often" (2), and "very often" (2) together.
* Significant fit residuals outside the range between −2.5 and +2.5 and refers to misfit of the item.

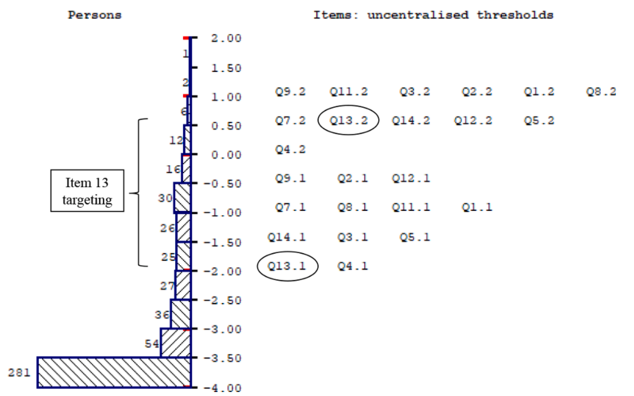

Fig. 1. Item-person threshold map of the final dataset OHIP-12 (after deleting item 6 and item 10). This figure displays the location of item threshold parameters and the distribution of person parameters along the latent dimension on the logit scale. Person-item maps help compare the range and position of the item thresholds distribution (right panel) to the range and position of the person parameter distribution (left panel). For each item (the right panel), the first symbol refers to the item/question number.
In contrast, the second digit refers to the threshold between the answer options after rescoring (two thresholds), including the three answer options "never" (0), "sometimes" (1) "often" (2). For example, item 13/ Q13 (Pain in the mouth) ranges from the threshold Q13.1 to Q13.2 and targets the highest number of persons (left panel). Items of a scale should ideally be located along the whole dimension to provide meaningful measures for all persons. The further down the patients are located, the better is their self-reported OHRQoL. Therefore, the continuum runs from high OHRQoL down to low OHRQoL.

least discriminating item as well as its age-related DIF; and the insufficient distinction of the five response options by the patients. Therefore, while the OHIP-14 met some of the measurement requirements, such as unidimensionality, it also revealed some deficiencies such as LD. Thus, our hypothesis was partially confirmed.

In an iterative process, we deleted two items. The newly proposed OHIP-12 achieved better model fit and higher reliability than the original OHIP-14, contributing to more efficient and clinically meaningful measurement.

Another limitation detected by our analysis is the targeting of the scale. The OHIP-14, as well as the newly derived OHIP-12, were off target for many respondents. Thus, the scale is not sensitive enough to assess the incipient decrease of OHRQoL in populations such as OA patients, who predominantly score at the better end of the scale. OA is a disease of higher frequency in older age [28], accordingly, most of our patients (83.9%) were more than 55 years old. Although we expected that older adults and/or people living with chronic conditions would show worse outcomes [24, 38, 39], the contrary was found within our study. Salde and Sanders reported "the paradox of better subjective oral health in older age", which could explain this diminishing impact of OHRQoL measured by the current OHIP [40]. Moreover, it could be challenging to measure the actual OHRQoL for this OA population because oral problems as a "non-major health problem" are not a significant concern given their overall health condition [41]. Therefore, those measurement-sensitive populations require adapted OHRQoL instrument to measure this diminishing perception of oral health problems accurately. Therefore, we recommend adding items that are more relevant to patients. For example, Item 13, "*Pain in the mouth*", cover a wide range of our respondents (**Fig. 1**). Such items addressing oral symptoms were found to improve the targeting of OHRQoL instruments [24, 41, 42]. Moreover, items among the most unlikely items to be affirmed by the respondents, such as items 2 (Taste is impaired), and 1 (Difficulties pronouncing certain words), should be explored further, and re-wording could also be considered. However, to explore this further, qualitative studies should be conducted [43].

The psychometric evaluation of the OHIP-14 was investigated in various studies, which reported Cronbach's alpha to be sufficient [7, 12]. However, the PSI as the Rasch equivalent of reliability turned out to be very low [15].This is due to the insufficient targeting and the lack of items at the better end of the scale, with 54.5% (281) of the patients reporting no impact on their OHRQoL. If we exclude extreme scores, the PSI was found satisfactory at the group level (0.7) (**Table 2**) [34].

The disordered thresholds could have occurred because of too many answer options or inadequate definitions of these, which could confuse the persons filling in the questionnaire [44]. Administering only three response options, such as "never", "occasionally", and "often", in future studies would test the amended OHIP scoring scheme empirically. Another possibility is to use the original answer options and accommodate disordered thresholds in the data analysis phase through the scoring scheme. Although similar results were reported in a psychometric study in Italy, to date, the proposed answer options have not been further investigated or applied when assessing OHRQoL using OHIP-14 [45].

Regarding our DIF finding by age-group in *item 5*, this was perceived as not meaningful DIF and should be reinvestigated in other samples to confirm or disconfirm DIF of people aged ≥76 years. OHIP-EDENT or OHIP-19 was explicitly developed to assess the OHRQoL of edentulous patients because it was considered more appropriate and representative to the edentulous population than OHIP-14 [6]. DIF was examined for the first time in OHIP-14 between dentate and edentulous patients. Our results didn't detect any DIF, indicating that OHIP-14 items are understood in the same way by both populations. Likewise, our results didn't show DIF in respondents with different self-reported severity of osteoarthritis.

Interestingly, LD was found among items in the same domain and between items of different domains. This observation suggests that the seven domains of OHIP-14 might be helpful for item generation, but they do not group items in an empirically reproducible way. Similar findings were found in other studies reporting redundancy in some of the OHIP items [46]. Regarding the OHIP-14 dimensional structure, studies have described different underlying dimensions of OHRQoL [10, 11]. However, the most recent ones have demonstrated that OHRQoL has four main conceptual components: Oral Function, Orofacial Pain, Orofacial Appearance, and Psychosocial Impact, which can be described by one overall score [8, 47, 48]. Similarly, the present study using the RM, which is perceived as a strict test for unidimensionality [49], did not reveal any indication of a multidimensional structure. Instead, the findings supported a unidimensional structure of OHIP-14, allowing for one summary measure characterizing OHRQoL [8, 9].

Our analysis provides evidence-based recommendations regarding the psychometric properties of OHIP-14 in an OA patients' sample. It also exposes an important insight of how the instrument is mistargeted with respect to the "general OA population". We recommend that further studies are conducted to improve the targeting of the scale for people with fewer OHRQoL problems. Accordingly, items could be replaced or added to cover the lower end of the scale based on qualitative methods; items from the longer original OHIP version (OHIP-49) could be considered. Consequently, for the "general OA population", we recommend using the current instrument when dental issues are suspected. As for the general screening, one should be aware of the limitations highlighted in our study. Moreover, for patients with health conditions such as OA, more investigations are needed to understand and confirm whether those patients do not perceive the OHRQoL problems because they are occupied with their other illness-related problems, or whether some of the OHIP items might not be functioning appropriately for this group of people.

The proposed OHIP-12 represents the set of OHIP-14 items that appears to be more appropriate for the measurement of OHRQoL in OA patients and potentially other chronic conditions or the elderly population. However, more studies using the RM are needed to calibrate OHIP-14 in different people, including the elderly, other chronic conditions, healthy, and primary distinct oral health conditions/seeking dental care.

5. Limitation

This study focused on diagnosing and adapting the OHIP-14 in its current state without addressing or adding new items to improve this instrument's essential qualities, such as targeting.

6. Conclusions

This study tackles for the first time the fundamental principles of measurement of OHIP-14 and proposes solutions with the newly suggested OHIP-12 as a more psychometrically accurate version of the instrument. The elimination of two items and a revised response scale featuring three options are appropriate and contribute to reducing response burden. The insufficient targeting of the scale in this group of patients indicates that the scale was not sensitive enough to assess the incipient decrease of OHRQoL in our sample. Hence, items addressing mild diminishment of OHRQoL should be added to the instrument.

Funding

None.

Author contributions

M. Omara, contributed to conception and design, contributed to data acquisition, analysis, and interpretation, drafted the manuscript, critically revised the manuscript; T. Salzberger, contributed to conception and design, contributed to data analysis and interpretation, critically revised the manuscript; M. Boecker, contributed to conception and design, contributed to data analysis and interpretation, critically revised the manuscript; K. Bekes, contributed to conception, critically revised the manuscript; G. Steiner, contributed to conception, contributed to data acquisition, critically revised the manuscript; V. Nell-Duxneuner, contributed to conception, contributed to data acquisition, critically revised the manuscript; V. Ritschl, contributed to conception, contributed to data acquisition, critically revised the manuscript; E. Mosor, contributed to conception, critically revised the manuscript; M. Kloppenburg, contributed to conception, critically revised the manuscript; J. Sautner, contributed to conception, critically revised the manuscript; B. Steinecker-Frohnwieser, contributed to conception, critically revised the manuscript; T. Stamm, contributed to conception and design, contributed to data acquisition, analysis, and interpretation, critically revised the manuscript. All authors gave their final approval and agree to be accountable for all aspects of the work.

Declaration of Competing Interest

None.

Acknowledgements

The authors would like to thank the participating patients and the clinical "BLOAR" team in all participating centres.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2021.103819.

References

[1] D. Bennadi, C. Reddy, Oral health related quality of life, J. Int. Soc. Prev. Commun. Dent. 3 (1) (2013) 1, https://doi.org/10.4103/2231-0762.115700.

[2] M. Omara, T. Stamm, M. Boecker, V. Ritschl, E. Mosor, T. Salzberger, C. Hirsch, K. Bekes, Rasch model of the Child Perceptions Questionnaire for oral health–related quality of life: a step forward toward accurate outcome measures, J. Am. Dental Ass. 150 (5) (2019) 352–361, https://doi.org/10.1016/j.adaj.2018.12.001, e7.

[3] G.D. Slade, A.J. Spencer, Development and evaluation of the Oral Health Impact Profile, Community Dent. Health. 11 (1) (1994) 3–11. https://pubmed.ncbi.nlm.nih.gov/8193981/.

[4] M.I. MacEntee, M. Brondani, Cross-cultural equivalence in translations of the oral health impact profile, Community Dent. Oral Epidemiol. 44 (2) (2016) 109–118, https://doi.org/10.1111/cdoe.12198.

[5] World Health Organization, International classification of impairments, disabilities, and handicaps: a manual of classification relating to the consequences of disease, published in accordance with resolution WHA29. 35 of the Twenty-ninth World Health Assembly, May 1976, (1980). https://apps.who.int/iris/handle/10665/41003.

[6] F. Allen, D. Locker, A modified short version of the oral health impact profile for assessing health-related quality of life in edentulous adults, Int. J. Prosthodont. 15 (5) (2002) 446–450. https://pubmed.ncbi.nlm.nih.gov/12375458/.

[7] G.D. Slade, Derivation and validation of a short-form oral health impact profile, Community Dent. Oral Epidemiol. 25 (4) (1997) 284–290, https://doi.org/10.1111/j.1600-0528.1997.tb00941.x.

[8] M.T. John, L. Feuerstahler, N. Waller, K. Baba, P. Larsson, A. Čelebić, D. Kende, K. Rener-Sitar, D.R. Reißmann, Confirmatory factor analysis of the oral health impact profile, J. Oral Rehabil. 41 (9) (2014) 644–652, https://doi.org/10.1111/joor.12191.

[9] C.M.d. Santos, B.H.d. Oliveira, P. Nadanovsky, J.B. Hilgert, R.K. Celeste, F. N. Hugo, Oral Health Impact Profile-14::¿ una escala unidimensional? Cadernos de Saúde Pública 29 (4) (2013) 749–757, https://doi.org/10.1590/s0102-311x2013000800012.

[10] J. Montero, M. Bravo, M.-.P. Vicente, M.-.P. Galindo, J.F. López, A. Albaladejo, Dimensional structure of the oral health-related quality of life in healthy Spanish workers, Health Qual. Life Outcomes 8 (1) (2010) 1–9, https://doi.org/10.1186/1477-7525-8-24.

[11] M.L. Zucoloto, J. Maroco, J.A.D.B. Campos, Psychometric properties of the oral health impact profile and new methodological approach, J. Dent. Res. 93 (7) (2014) 645–650, https://doi.org/10.1177/0022034514533798.

[12] M.T. John, D.L. Miglioretti, L. LeResche, T.D. Koepsell, P. Hujoel, W. Micheelis, German short forms of the Oral Health Impact Profile, Community Dent. Oral Epidemiol. 34 (4) (2006) 277–288, https://doi.org/10.1111/j.1600-0528.2006.00279.x.

[13] D. Andrich, A general form of Rasch's extended logistic model for partial credit scoring, Applied Measurement in Education 1 (4) (1988) 363–378, https://doi.org/10.1207/s15324818ame0104_7.

[14] T. Bond, C.M. Fox, Applying the Rasch Model: fundamental Measurement in the Human Sciences, Third Edition (2015), https://doi.org/10.4324/9781410614575.

[15] J.F. Pallant, A. Tennant, An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS), Br. J. Clin. Psychol. 46 (1) (2007) 1–18, https://doi.org/10.1348/014466506X96931.

[16] P. Panayides, C. Robinson, P. Tymms, The assessment revolution that has passed England by: rasch measurement, Br Educ. Res. J. 36 (4) (2010) 611–626, https://doi.org/10.1080/01411920903018182.

[17] G. Rasch, An individual-centered approach to item analysis with two categories of answers, Psychological measurement theory, proceedings of the NUFFIC international session of science at" Het Oude Hof" The Hague, July (1966) 14–Z8. https://lies.mat.uc.cl/wp-content/uploads/2020/12/memo19642.pdf.

[18] G. Rasch, Probabilistic models for some intelligence and attainment tests, (1993). https://eric.ed.gov/?id=ED419814.

[19] E.V. Smith Jr., Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals, J. Appl. Meas. 3 (2) (2002) 205–231. https://pubmed.ncbi.nlm.nih.gov/12011501/.

[20] K.B. Christensen, G. Makransky, M. Horton, Critical values for Yen's Q 3: identification of local dependence in the Rasch model using residual correlations, Appl. Psychol. Meas. 41 (3) (2017) 178–194, https://doi.org/10.1177/0146621616677520.

[21] G. Tutz, G. Schauberger, A penalty approach to differential item functioning in Rasch models, Psychometrika 80 (1) (2015) 21–43, https://doi.org/10.1007/s11336-013-9377-6.

[22] J. Petrillo, S.J. Cano, L.D. McLeod, C.D. Coon, Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples, Value Health 18 (1) (2015) 25–34, https://doi.org/10.1016/j.jval.2014.10.005.

[23] T. Stamm, M. Mathis, D. Aletaha, M. Kloppenburg, K. Machold, J. Smolen, Mapping hand functioning in hand osteoarthritis: comparing self-report instruments with a comprehensive hand function test, Arthritis Rheum. 57 (7) (2007) 1230–1237, https://doi.org/10.1002/art.22989.

[24] J.L. Kelsey, I.B. Lamster, Influence of musculoskeletal conditions on oral health among older adults, Am. J. Public Health 98 (7) (2008) 1177–1183, https://doi.org/10.2105/AJPH.2007.129429.

[25] X. Wang, J. Zhang, Y. Gan, Y. Zhou, Current understanding of pathogenesis and treatment of TMJ osteoarthritis, J. Dent. Res. 94 (5) (2015) 666–673, https://doi.org/10.1177/0022034515574770.

[26] J.W. Kim, M.K. Chung, J. Lee, S.K. Kwok, W.U. Kim, S.H. Park, J.H. Ju, Association of periodontitis with radiographic knee osteoarthritis, J. Periodontol. 91 (3) (2020) 369–376, https://doi.org/10.1002/JPER.19-0068.

[27] S. Safiri, A.-.A. Kolahi, E. Smith, C. Hill, D. Bettampadi, M.A. Mansournia, D. Hoy, A. Ashrafi-Asgarabad, M. Sepidarkish, A. Almasi-Hashiani, Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the Global

Burden of Disease Study 2017, Ann. Rheum. Dis. 79 (6) (2020) 819–828, https://doi.org/10.1136/annrheumdis-2019-216515.
[28] D.J. Hunter, S. Bierma-Zeinstra, Osteoarthritis, Lancet 393 (10182) (2019) 1745–1759, https://doi.org/10.1016/S0140-6736(19)30417-9.
[29] J.P. Vandenbroucke, E. von Elm, D.G. Altman, P.C. Gøtzsche, C.D. Mulrow, S. J. Pocock, C. Poole, J.J. Schlesselman, M. Egger, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration, Epidemiology 18 (6) (2007) 805–835, https://doi.org/10.1371/journal.pmed.0040297.
[30] Johns Hopkins Arthritis Center, American College of Rheumatology Diagnostic Guidelines. https://www.hopkinsarthritis.org/physician-corner/education/arthritis-education-diagnostic-guidelines/. (Accessed 29. 07 2020).
[31] P. Hagell, A. Westergren, Sample size and statistical conclusions from tests of fit to the rasch model according to the rasch unidimensional measurement model (rumm) program in health outcome measurement, J. Appl. Meas. 17 (4) (2016) 416–431. https://pubmed.ncbi.nlm.nih.gov/28009589/.
[32] A.B. Sadosky, A.G. Bushmakin, J.C. Cappelleri, D.R. Lionberger, Relationship between patient-reported disease severity in osteoarthritis and self-reported pain, function and work productivity, Arthritis Res. Ther. 12 (4) (2010) R162, https://doi.org/10.1186/ar3121.
[33] A. Tennant, J. Pallant, Unidimensionality matters!(A Tale of Two Smiths?), Rasch Measur. Trans. 20 (1) (2006) 1048–1051. https://www.rasch.org/rmt/rmt201c.htm.
[34] A. Tennant, P.G. Conaghan, The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 57 (8) (2007) 1358–1362, https://doi.org/10.1002/art.23108.
[35] S. UNESCO, United Nations Educational, C. Organization, International Standard Classification of Education, ISCED 1997 (2003), https://doi.org/10.1007/978-1-4419-9186-7_10.
[36] K.B. Christensen, K. Thorborg, P. Hölmich, M.B. Clausen, Rasch validation of the Danish version of the shoulder pain and disability index (SPADI) in patients with rotator cuff-related disorders, Qual. Life Res. 28 (3) (2019) 795–800, https://doi.org/10.1007/s11136-018-2052-8.
[37] P.L. Rumm Laboratry, RUMM2030. Released in January 2010. License re-structure from March 2012, (2012).
[38] C. Dörfer, C. Benz, J. Aida, G. Campard, The relationship of oral health with general health and NCDs: a brief review, Int. Dent. J. 67 (2017) 14–18, https://doi.org/10.1111/idj.12360.
[39] G. Schmalz, S. Patschan, D. Patschan, D. Ziebolz, Oral-health-related quality of life in adult patients with rheumatic diseases—a systematic review, J. Clin. Med. 9 (4) (2020) 1172, https://doi.org/10.3390/jcm9041172.
[40] G.D. Slade, A.E. Sanders, The paradox of better subjective oral health in older age, J. Dent. Res. 90 (11) (2011) 1279–1285, https://doi.org/10.1177/0022034511421931.
[41] E. Rodakowska, M. Wilczyńska-Borawska, J. Fryc, J. Baginska, B. Naumnik, Oral health-related quality of life in patients undergoing chronic hemodialysis, Patient Prefer Adherence 12 (2018) 955, https://doi.org/10.2147/PPA.S161638.
[42] T.A. Stamm, M. Omara, S.R. Bakerc, L.F. Page, W.M. Thomson, P.E. Benson, T. Broomhead, F. Aguilar-Diaz, L. Do, B.J. Gibson, Rasch model of the child perceptions questionnaire in multi-country data, J. Dent. 93 (2020), 103267, https://doi.org/10.1016/j.jdent.2019.103267.
[43] D. Locker, F. Allen, What do measures of 'oral health-related quality of life'measure? Community Dent. Oral Epidemiol. 35 (6) (2007) 401–411, https://doi.org/10.1111/j.1600-0528.2007.00418.x.
[44] D. Andrich, A rating formulation for ordered response categories, Psychometrika 43 (4) (1978) 561–573, https://doi.org/10.1007/BF02293814.
[45] M. Franchignoni, A. Giordano, E. Brigatti, M. Migliario, L. Levrini, G. Ferriero, [Psychometric properties of the Italian version of the reduced form of the Oral Health Impact Profile (OHIP-14)], G. Ital. Med. Lav. Ergon. 32 (3 Suppl B) (2010) B71–B78. https://pubmed.ncbi.nlm.nih.gov/21302527/.
[46] S. Baker, B. Gibson, D. Locker, Is the oral health impact profile measuring up? Investigating the scale's construct validity using structural equation modelling, Community Dent. Oral Epidemiol. 36 (6) (2008) 532–541, https://doi.org/10.1111/j.1600-0528.2008.00440.x.
[47] M.T. John, S. Sekulić, K. Bekes, M.H. Al-Harthy, A. Michelotti, D.R. Reissmann, J. Nikolovska, S. Sanivarapu, F.B. Lawal, T. List, Why patients visit dentists–A study in all WHO regions, J. Evid. Based Dental Practice (2020), 101459, https://doi.org/10.1016/j.jebdp.2020.101459.
[48] M.T. John, B. Häggman-Henrikson, S. Sekulic, T. Stamm, I. Oghli, O. Schierz, T. List, K. Baba, K. Bekes, A. van Wijk, Mapping Oral Disease Impact with a Common Metric (MOM)—Project summary and recommendations, J. Oral Rehabil. 48 (3) (2021) 305–307, https://doi.org/10.1111/joor.13133.
[49] P. Hagell, Testing rating scale unidimensionality using the principal component analysis (PCA)/*t*-test protocol with the Rasch model: the primacy of theory over statistics, Open J Stat 4 (6) (2014) 456–465, https://doi.org/10.4236/ojs.2014.46044.