Community Dent Oral Epidemiol 2006; 34: 277–88

Community Dentistry and Oral Epidemiology

German short forms of the Oral Health Impact Profile

Mike T. John[1,2], Diana L. Miglioretti[3], Linda LeResche[2], Thomas D. Koepsell[4], Philippe Hujoel[5] and Wolfgang Micheelis[6]

[1]Department of Prosthodontics and Materials Science, School of Dentistry, University of Leipzig, Germany, [2]Department of Oral Medicine, School of Dentistry, University of Washington, Seattle, WA, USA, [3]Center for Health Studies, Group Health Cooperative, and Department of Biostatistics, University of Washington, Seattle, WA, USA, [4]Department of Epidemiology and Department of Health Services, School of Public Health and Community Medicine, University of Washington, Seattle, WA, USA, [5]Department of Dental Public Health Sciences, School of Dentistry, University of Washington, Seattle, WA, USA, [6]Institute of German Dentists, Universitätsstr, Köln, Germany

John MT, Miglioretti DL, LeResche L, Koepsell TD, Hujoel P, Micheelis W. German short forms of the Oral Health Impact Profile. Community Dent Oral Epidemiol 2006; 34: 277–88. © Blackwell Munksgaard, 2006

Abstract – *Objectives:* We report the development and psychometric evaluation of short forms of the Oral Health Impact Profile German version (OHIP-G) - an instrument to assess oral health-related quality of life (OHRQoL). *Methods:* A five-item short form was developed using best subset regression in 2050 subjects from a national survey. Two 14-item versions were derived from English-language short forms and a 21-item version from previous factor analytic work. A second sample from the general population ($n = 163$) and a sample of clinical patients with temporomandibular disorders (TMD; $n = 175$) were used to investigate validity and internal consistency. Test-retest reliability was evaluated in 30 prosthodontic patients before treatment. Responsiveness was assessed in 67 patients treated for their TMD pain. *Results:* Associations between short form summary scores and self-report of oral health and four oral disorders in the general population and in TMD patients were interpreted as support for convergent/groups validity. The instruments' responsiveness (effect measures of 0.55–0.98), test-retest reliability (intraclass correlation coefficients: 0.72–0.87), and internal consistency (Cronbach's alpha: 0.65–0.92) were sufficient. *Conclusions:* Sufficient discriminative and evaluative psychometric properties of short forms of the OHIP-G make the instruments suitable to assess OHRQoL in cross-sectional as well as longitudinal studies.

Key words: health-related quality of life; oral health-related quality of life; questionnaire; reliability; short instruments; validity

Dr Mike T. John, Department of Prosthodontics and Materials Science, University Leipzig, Nürnberger Str. 57, 04103 Leipzig, Germany
Tel: 49 341 9721 363
Fax: 49 341 9721 329
e-mail: mike.john@medizin.uni-leipzig.de

Submitted 5 January 2005;
accepted 1 September 2005

Recently a German version of the Oral Health Impact Profile (OHIP-G) was developed (1) to measure oral health-related quality of life (OHRQoL). Both the German and the English-language version (2) are sophisticated instruments. However, the time required to ask and answer the 53 questions in the German version or 49 items in the English-language form sets restrictions to their application in surveys and health care settings. In 1996, a review of OHRQoL instruments presented the OHIP as the longest instrument among 11 oral health-specific instruments (3). Five questionnaires had between three and 14 items. To the extent that these instruments also perform well compared with the OHIP-G, this may indicate that in Germany the construct OHRQoL may be adequately described by fewer than 53 items.

Indeed, examination of OHIP-G dimensional structure (4) showed a substantial amount of correlation among items. This 'information redundancy' indicates a potential for grouping highly correlated items into latent variables or constructs (dimensions) but also allows for item reduction (shortening the instrument). Not all items may need to be measured. Excluding the four German-specific items and three items relevant only for subjects with dentures, we found that only 21 of 46 OHIP items loaded on four identified OHRQoL dimensions (orofacial pain, oral functions, psychosocial impact, appearance). Although results are based on an exploratory factor analysis and need to be supported by confirmatory factor analysis and validation, these results point in the direction of possible item reduction and therefore abbreviating the OHIP-G.

Desired characteristics of a short German Oral Health Impact Profile

Health and health status measures are context specific. They need to be cross-culturally adapted (5, 6). Increasing international collaboration and globalization in medical research demand the existence of cross-culturally compatible instruments to assess (oral) health-related quality of life (7). International compatibility is a priority for a new instrument. The use of the 14-item short Oral Health Impact Profile (OHIP-14, originally developed in Australia) in other English-speaking countries and its close relationship to other OHRQoL measures (8, 9), its apparently good psychometric properties (10), and the existence of a long German instrument (1) make this questionnaire the first choice for the development of a short German OHRQoL instrument.

In our opinion, the advantages of using a German version of the OHIP-14 for different research and patient-oriented applications could be outweighed if another instrument which is much shorter than 14 items could be used to assess the same construct. Such a very short instrument could be routinely used in nondental settings. So far, none of the general health status measures includes oral health specific components (11).

A second possible reason not to simply translate the OHIP-14 has to do with OHRQoL dimensions. Previously, we showed that the dimensional structure of the long English-language OHIP was not optimal for the adult German population (4). However, we presented a smaller set of 21 items that measured orofacial pain, oral functions, psychosocial impact, and appearance as possible dimensions for OHRQoL. Describing OHRQoL dimensions with a questionnaire substantially shorter than 49 or 53 items would be worthwhile. Because subjective oral health indicators have many political, theoretical, and practical applications (12), a short measure describing the entire construct and components of OHRQoL should be even more useful because of the deeper insight into relevant aspects of patients perceptions. The aim of this study was to develop one or more German short forms of the OHIP and to submit them to psychometric evaluation.

Materials and methods

Short German versions of the OHIP were developed in two steps. First, several instruments were developed by adaptation of the English version and by selecting items based on statistical performance of the abbreviated scale. The second step involved the assessment of the psychometric properties (reliability, validity, and responsiveness) of the new instruments. Two samples of the general population and three patient samples attending the Department of Prosthodontics, Martin Luther University Halle-Wittenberg, Germany were used (Table 1).

Development of instruments

English-language instruments

We selected two available short versions of the Oral Health Impact Profile each with 14 items to be crossculturally adapted. The first was published by Slade in 1997 (10). The second short English-language OHIP was published by Locker and Allen in 2002 (13).

The German items equivalent to the items in the two 14-item English-language OHIP versions were selected from the available item pool created for the cross-culturally adapted long OHIP-G (1).

New German-specific instruments

In addition to the two existing English-language instruments, two German instruments were developed. First, based on our previous work, where factor analytic methods were applied to identify dimensions of OHRQoL, 21 items were identified which loaded on four OHRQoL dimensions (oral Functions, orofacial pain, psychosocial impact, appearance; 4). Therefore, we used these 21 items as a short form (OHIP-G21).

A second new instrument was desired, which could capture a large proportion of information in the OHIP summary score with the minimum number of items. We selected as a criterion that at least 90% of the variance (adjusted R^2) of the OHIP-G summary score (sum of all item responses) should be explained by the items selected. Using 2050 subjects from a national survey (Table 1, sample A), ordinary least squares (OLS) best subset regression was used to select the items. The subset regression searched through all combinations of items to select a combination of items that performs 'best' according to a defined statistical criterion (in our case to explain at least 90% OHIP-G summary score variance). Therefore, *all* possible statistical models were evaluated, which is an advantage over stepwise selection approaches that include and exclude items according to certain criteria and do *not* evaluate

Table 1. Overview of sampling strategies, data collection methods, and sampled populations by age, gender, and research purpose

Sample	Sample type	Data collection	*n*	Age [mean (SD)]	Age range	% women	Type of investigation
(A) German speaking individuals living in private households of the Federal Republic of Germany aged 16–79 years (response proportion: 60%) selected with a multistage sampling technique from subjects registered at the community population register office	Random	Interview	2050	43.3 (16.2)	16–79	52	Development of shortest possible instrument
(B) Individuals in the metropolitan area of Halle/Saale (response proportion: 54%) selected with a stratified sample from subjects registered at the population register office	Random	Questionnaire[b]	163	38.3 (11.3)	20–60	67	Construct validity and internal consistency
(C) Patients with temporomandibular disorder (Department of Prosthodontics, MLU[a])	Consecutive	Interview	175	39.2 (15.8)	18–85	78	Construct validity and internal consistency
(D) Patients with temporomandibular disorder pain (Department of Prosthodontics, MLU[a])	Consecutive	Interview	67	41.3 (15.6)	19–85	72	Responsiveness
(E) Patients with prosthodontic treatment need (Department of Prosthodontics, MLU[a])	Convenience	Questionnaire[b]	30	50.7 (21.1)	18–85	53	Test-retest reliability

[a]Martin Luther University Halle-Wittenberg.
[b]Interviewer-supervised, self-administered questionnaire.

all models. Best subset regression started with single variable models and went up to eight variables. The *best* 10 models for a specified number of items were compared. Among those 10 models, selection of the final model was based on both statistical properties and coverage of subject matter and interpretability.

The ability to predict the long OHIP-G summary score was validated in a regional random sample of the general population and in a convenience sample of patients with temporomandibular disorders (TMD; Table 1, samples B and C).

Evaluation of psychometric properties

Content validity

Content validity was sought by using the item pool of the long OHIP-G. All short forms were compared with the seven OHRQoL dimensions (functional limitation, physical pain, psychological discomfort, physical disability, psychological disability, social disability, and handicap) incorporated into the design of the long English-language version (2) and the four dimensions (orofacial pain, oral functions, psychosocial impact, and appearance) that emerged from our previous factor analytic work on the German version (4).

Construct validity

Construct validity was examined in a regional random sample of the general population and a convenience sample of TMD patients (Table 1, samples B and C). Summary scores calculated as the simple sum of all 49 item frequencies contained in the English-language OHIP (the four German-specific items were omitted to maintain international comparability) represented the construct OHRQoL. Higher scores imply poorer OHRQoL because the OHIP index measures the frequency of problems. Construct validity was evaluated by examining the association between self-reported oral health (very good, good, fair, poor) or self-report of several oral conditions and the short form OHIP-G summary scores. These oral conditions were defined as follows:

- TMD pain in the last month according to a question in the Research Diagnostic Criteria for Temporomandibular Disorders (14),
- Burning mouth sensations in the last 6 months according to a question from a US-national survey (15),
- Self-report of halitosis (never, hardly ever, sometimes, often), and
- Oral habits (yes/no) defined as biting on nails, tongue, lip, cheek, or objects.

It was expected that subjects with no TMD pain, no burning mouth sensations, less frequently reported bad breath, and better self-reported oral health would have lower short form OHIP-G scores, i.e. the mean of OHIP scores for these subjects should be lower compared with subjects without or with lesser extend of these conditions. For the fourth condition, oral habits, a substantial influence on OHRQoL was not expected *a priori*. A clinically relevant and a statistically significant association between this condition and short-form OHIP-Gs should therefore be absent.

Spearman rank correlations were calculated to examine the associations between global rating of oral health or frequency of halitosis and the short form OHIP-G summary scores. Point-biserial correlations were calculated to examine the associations between TMD pain, oral habits, burning mouth sensations, and the summary scores.

Responsiveness

Patients with TMD pain (Table 1, sample D) were chosen for the assessment of the instruments' responsiveness because orofacial pain, like other chronic pain conditions, has a major impact on the patient's quality of life (16). A randomized controlled trial in our TMD pain patient population showed a clinically relevant and statistically significant effect of a dental treatment over the time period of 1 month (17). Based on the close relationship between pain and quality of life (18) we hypothesized that OHRQoL would improve over this period, too. Using treated TMD patients would therefore allow us to assess responsiveness of the short instruments. The short form OHIP-G summary score change from baseline to follow up was tested using the paired *t*-test. A measure of responsiveness (standardized effect size) was calculated as: (Mean baseline OHIP score − follow up OHIP score)/standard deviation of baseline OHIP score according to Allen et al. (19). A second, similar measure, the standardized response mean, was calculated as: (Mean baseline OHIP score − follow up OHIP score)/(standard deviation of baseline OHIP score − follow up OHIP score).

Reliability

Test-retest reliability was assessed in a convenience sample of patients with a prosthodontic treatment

need (Table 1, sample E) using a time interval of 2 weeks between the administration of the two questionnaires. Intraclass correlation coefficients (ICC) were calculated for all short forms according to Shrout's and Fleiss's ICC(2,1) (20). In addition, the method of Bland and Altman (21) was used to detect systematic differences between the two measures and to quantify test-retest differences. It involved the computation of the standard deviation of the differences between the measures at time points 1 and 2. 'Limits of agreement' around the mean difference were calculated as 1.96 times the standard deviation of the differences. Hence, this statistic represents the test-retest differences expected for 95% of the individuals in the sample. In addition, a confidence interval for the mean of the differences was computed. If it excludes zero, it indicates a statistically significant difference between the measures at time points 1 and 2.

Internal consistency was measured in a regional random sample of the general population and in a convenience sample of TMD patients (Table 1, samples B and C) using Cronbach's alpha (22) and interitem correlation.

Missing data

Missing items would compromise the calculation of summary scores. In samples A–D (Table 1) data were missing. Among the 2050 subjects of the national survey (sample A), 261 answers were missing in 174 subjects after dropping 24 subjects according to previously specified criteria (23). Five answers were missing in the regional random sample of the general population (sample B), two values in the convenience sample of TMD patients (sample C), and 10 answers in the sample of consecutive TMD pain patients (sample D). Missing answers were imputed using regression imputation [see (23) for details].

All calculations except best subset regression were carried out with the statistical package STATA (Stata Statistical Software Release 7, 1999; StataCorp., College Station, TX, USA). Best subset regression was performed using the statistical package SAS (SAS/STAT Statistical Software, Release 6.12, 1996; SAS Institute, Inc., Cary, NC, USA).

Results

Five independent samples containing together 2485 subjects and covering an age range between 16 and 85 represented a wide range of the adult population (Table 1).

Development of an efficient OHIP-G short form

In sample A best subset regression searched through all statistical models containing one to eight items (Fig. 1) to identify an item combination that explained the information in the long OHIP-G summary score efficiently. The best one-item model already explained 60% of the variance of the summary score. Five-item models exceeded the 90% adjusted R^2 criterion. Rapidly decreasing amounts of additional variance were explained after that point.

More than 100 of the five-item models exceeded the 90% adjusted R^2 criterion. The top 10 models were indistinguishable according to that criterion (Table 2), because the model with the lowest adjusted R^2 differed by only 0.0064 from the maximum model. Based on subject matter, model 9 was selected. This model contained the most characteristic item for the dimension appearance ('felt uncomfortable about appearance'), which was considered a better representation of the dimension appearance than 'worried'. Additionally, 'difficulty chewing' was more prevalent in the sample than 'trouble pronouncing any words' and 'painful aching' occurred more frequently than 'sore jaw' (23). More prevalent items were considered superior to less prevalent items.

The five selected items performed well in predicting the long OHIP-G summary score in a different sample of the general population (sample

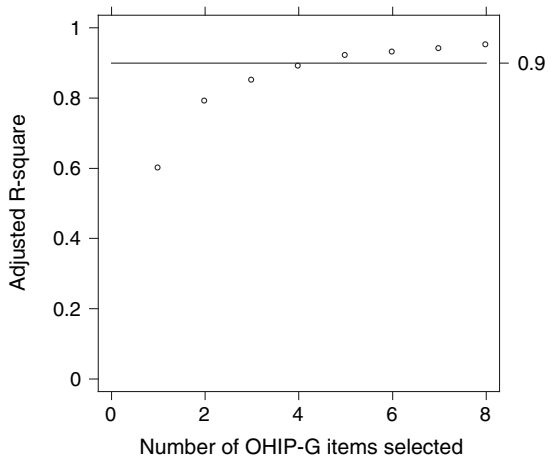

Fig. 1. Relationship between number of items selected by best subset regression and explained variance of the independent variable (OHIP-G49 summary score) assessed in sample A.

Table 2. Top 10 regression models explaining OHIP-G49 variance with five items and their assignment to OHRQoL dimensions (final model: no. 9)

		Item number and assignment to dimensions from exploratory factor analysis				
No	Adjusted R^2	Function	Pain	Appearance	No assignment[a]	Psychosocial impact
1	0.9152	2	10	19	23	43
2	0.9131	2	10	19	26	43
3	0.9127	1		19	26, 27	43
4	0.9113	2	10	22	23	43
5	0.9108	1	11	19	26	43
6	0.9107	2	11	19	26	43
7	0.9103	2	11	19	23	43
8	0.9100	1		19	26, 29	43
9	0.9099	1	10	22	26	43
10	0.9088	1		19	26, 32	43

[a]All 46 OHIP items (excluding the German-specific items and items referring to subjects with dentures) used for the explanatory factor analysis could be assigned to one of four dimensions ($n = 21$ items) or remained without dimension assignment ($n = 25$ items).
Item nos: 1, difficulty chewing foods; 2, trouble pronouncing words; 10, painful aching; 11, sore jaw; 19, worried; 22, felt uncomfortable about appearance; 23, felt tense; 26, felt less flavor in food; 27, unable to brush teeth properly; 29, unsatisfactory diet; 32, interrupted meals; 43, difficulty doing usual jobs.

B) and in a sample of TMD patients (sample C). The adjusted R^2 was 0.88 for the general population and 0.82 for TMD patients. In comparison, the 14-item and the 21-item OHIP-G versions explained between 91% and 96% of the variance, respectively.

Psychometric properties of the German short forms

Content validity

The OHIP dimensions were well-represented in short instruments (Table 3). Except for the dimension *physical disability*, all English-language scales were represented in the OHIP-G21. OHIP-G14A did not contain the German dimension *appearance*. Given that short instruments necessarily lose validity because they have fewer items, content validity was considered sufficient.

Construct validity

In samples B and C, all observed associations between self-report of oral conditions and OHRQoL followed the hypothesized direction or were absent when hypothesized to be absent (Tables 4 and 5). Overall self-rating of oral health and three conditions with hypothesized effects on OHRQoL were associated with the summary scores. Similar magnitudes of the correlations and levels of statistical significance were observed for all short forms (Table 6). Coefficients were positive where a relationship between the presence of a condition and impaired OHRQoL was expected. For report of oral habits where no association was expected coefficients were around zero. We consider these patterns to be consistent with our prior hypotheses.

Responsiveness

In sample D, all short forms were sensitive to a change of the construct OHRQoL (responsiveness) indicated by clinically relevant and statistically significant decreases in summary scores over the 1 month treatment period (Table 7). Ranking of effect size magnitude was similar when standardized effect sizes were compared with standardized response means. The effect size was largest for the OHIP-G5.

Reliability

In sample E, three short OHIPs with 14 and 21 items reached excellent test-retest reliability [ICC > 0.75, (24)]; the shortest instrument (OHIP-G5) reached almost this level with 0.72 (Table 8). Except for the OHIP-G21, which was the longest instrument and had therefore the largest chance to detect differences, no systematic differences were observed between the measures at the first time point and at follow-up. Limits of agreement showed considerable variability in the differences between the two measures.

Cronbach's alpha, assessed in samples B and C, showed the same pattern (Table 8). The shortest instrument had the smallest values and the longest instrument the largest. The two instruments with 14 items had almost identical values. Average interitem correlations were similar for all OHIPs.

Table 3. Content validity of short German OHIPs – number of instruments' items contained in original English-language OHIP and OHIP-G's dimension

	Item number in original English-language OHIP (2)			
	OHIP-G5	OHIP-G21	OHIP-G14a	OHIP-G14b
English-language OHIP dimension (number of items contained in dimension)				
Functional limitation (9)	1	1,2,3,4	2,6	1,7
Physical pain (9)	10	10,11,13,14,15,17	10,16	13,17
Psychological discomfort (5)	22	19,22	20,23	19,21
Physical disability (9)	26		29,32	24,28
Psychological disability (6)		36,37,38	35,38	34,36
Social disability (5)		39,40,42,43	42,43	40,42
Handicap (6)	43	48,49	47,48	45,47
OHIP-G dimension (number of items contained in dimension)				
Psychosocial impact (9)	43	36,37,38,39,40,42,43,48,49	38,42,43,48	36,40,42
Orofacial pain (6)	10	10,11,13,14,15,17	10	13,17
Oral functions (3)	1	1,2,4	2	1
Appearance (3)	22	3,19,22		19
Remaining items	26		6,16,20,23,29,32,35,47	7,21,24,28,34,45,47

Table 4. Construct validity for the five- and 21-item OHIP-Gs: associations between self-rating of oral health and self-report of four oral conditions in the general population (sample B) and temporomandibular disorder (TMD) patients (sample C)

	General population ($n = 163$)			TMD patients ($n = 175$)[a]		
		Mean			Mean	
Variable	n	OHIP-G5	OHIP-G21	n	OHIP-G5	OHIP-G21
Self-rating of oral health						
Very good	13	0.3	2.2	14	3.7	6.4
Good	106	0.8	4.8	109	6.0	16.9
Fair	39	3.1	14.0	39	9.2	27.2
Poor	5	7.2	26.6	10	9.1	28.6
TMD pain						
No	146	1.3	6.6	42	4.1	11.2
Yes	17	3.7	15.1	130	7.6	21.6
Burning mouth sensations						
No	161	1.4	7.3	146	6.5	17.4
Yes	2	7.0	22.0	25	8.0	28.5
Halitosis						
Never	61	0.7	4.0	75	5.9	16.0
Hardly ever	44	1.4	7.8	20	5.4	15.7
Sometimes	50	2.5	11.3	42	7.7	24.2
Often	8	2.8	8.4	19	7.4	20.7
Report of oral habits						
No	98	1.4	6.6	119	7.0	19.2
Yes	65	1.7	8.8	51	6.1	19.0

[a]Missing data for TMD patients: three subjects for self-rating of oral health and TMD pain, four subjects for burning mouth sensations, 19 subjects for self-report of halitosis and five subjects for report of oral habits.

Measures for internal consistency were higher for the general population sample compared with the TMD patients.

All results from test-retest reliability and internal consistency analyses were consistent which each other and therefore provided evidence for sufficient reliability of the OHIP-G short forms.

Discussion

Several German short forms of the Oral Health Impact Profile appear to be valid and reliable instruments. Different forms may be useful for different purposes. A short version containing only five items might be one of the most parsimonious

Table 5. Construct validity for the two 14-item OHIP-Gs: associations between self-rating of oral health and self-report of four oral conditions in the general population (sample B) and temporomandibular disorder (TMD) patients (sample C)

	General population (n = 163)			TMD patients (n = 175)[a]		
		Mean			Mean	
Variable	n	OHIP-G14A	OHIP-G14B	n	OHIP-G14A	OHIP-G14B
Self-rating of oral health						
Very good	13	0.7	2.0	14	5.9	6.4
Good	106	2.2	4.5	109	12.7	13.5
Fair	39	7.6	10.5	39	19.5	19.1
Poor	5	17.6	19.8	10	20.6	22.0
TMD pain						
No	146	3.2	5.6	42	8.9	9.6
Yes	17	10.1	11.3	130	15.8	16.4
Burning mouth sensations						
No	161	3.7	6.1	146	13.4	13.9
Yes	2	17.0	12.5	25	18.1	19.4
Halitosis						
Never	61	1.9	3.7	75	11.4	12.3
Hardly ever	44	3.8	6.6	20	11.4	12.4
Sometimes	50	6.2	8.7	42	17.9	18.0
Often	8	5.0	7.9	19	16.2	16.7
Report of oral habits						
No	98	3.3	5.5	119	14.4	14.8
Yes	65	4.7	7.3	51	13.6	14.5

[a]Missing data for TMD patients: three subjects for self-rating of oral health and TMD pain, four subjects for burning mouth sensations, 19 subjects for self-report of halitosis and five subjects for report of oral habits.

Table 6. Construct validity for four short OHIPs: magnitude of correlation coefficients and level of statistical significance for associations between self-rating of oral health and self-report of four oral conditions and OHIP summary scores in the general population (sample B) and temporomandibular disorder (TMD) patients (sample C)

		Correlation coefficients/statistical significance			
Variable	Sample	OHIP-G5	OHIP-G21	OHIP-G14a	OHIP-G14b
Self-rating of oral health[a]	General population	0.50***	0.54***	0.44***	0.51***
	TMD patient	0.40***	0.45***	0.39***	0.39***
TMD pain[b]	General population	0.29***	0.28***	0.33***	0.25**
	TMD patient	0.38***	0.33***	0.30***	0.31***
Burning mouth sensations[b]	General population				
	TMD patient	0.13	0.29***	0.17*	0.20**
Halitosis[a]	General population	0.31***	0.31***	0.24**	0.34***
	TMD patient	0.19*	0.20*	0.27***	0.21**
Report of oral habits[b]	General population	0.07	0.12	0.11	0.13
	TMD patient	−0.11	0.00	−0.04	−0.01

*$0.01 \le P < 0.05$; **$0.001 \le P < 0.01$; ***$P < 0.001$.
[a]Spearman rank correlation coefficient.
[b]Point-biserial correlation coefficient.

Table 7. Responsiveness for the four short OHIP's assessed in sample D

Responsiveness measure	OHIP-G5	OHIP-G21	OHIP-G14a	OHIP-G14b
Mean baseline score/mean follow-up score	7.5/3.9	19.8/12.8	15/8.7	15.6/10.0
P-value for difference	<0.001	<0.001	<0.001	<0.001
Summary score range at baseline	1–20	2–64	2–49	0–39
Standardized effect size	0.95	0.55	0.61	0.61
Standardized response mean	0.98	0.68	0.71	0.69

Table 8. Test-retest reliability measured by ICC's and the Bland–Altman Method assessed in sample E and internal consistency characterized by Cronbach's alpha and interitem correlation for all four short OHIP's (sample B)

Reliability measure	OHIP-G5	OHIP-G21	OHIP-G14a	OHIP-G14b
ICC	0.72	0.79	0.83	0.87
Mean difference (95% CI)	0.3 (−0.3 to 1.0)	2.1 (0.6 to 3.6)*	1.1 (−0.3 to 2.4)	1.0 (−0.1 to 2.2)
Limits of agreement	−3.2 to 3.9	−6.0 to 10.2	−6.1 to 8.2	−5.2 to 7.3
Cronbach's alpha				
General population	0.76	0.92	0.90	0.90
TMD patients	0.65	0.91	0.87	0.86
Average interitem correlation				
General population	0.39	0.37	0.40	0.39
TMD patients	0.31	0.32	0.33	0.31

*$P < 0.01$.

ways to characterize the construct OHRQoL. The 14-item version containing the identical items as the original English-language abbreviated form (10) would serve well in studies where internationally comparable results are the priority. We did not find a superior performance of the later-developed English-language short OHIP in the psychometric evaluation. Validity, reliability, and responsiveness for both instruments were almost identical for both the general population sample and the TMD patients. Although we expect a reduction of floor effects of the summary scores (i.e. fewer very low scores) for the later-developed OHIP (13), we do not consider this advantage important enough to recommend the later-developed version for routine use in German settings instead of the original one, which is an established instrument.

A longer 21-item version is not only able to describe the overall construct OHRQoL, but also has the potential to characterize dimensions such as orofacial pain, oral functions, psychosocial impact, and appearance. The good discriminative and evaluative psychometric properties of all instruments make them suitable for cross-sectional as well as longitudinal studies.

Psychometric properties

Construct validity and responsiveness of the long OHIP-G (1) were similar to the long English-language OHIP which is well established (2). Similar findings for oral health values in a cross-cultural study further support OHIP's validity (25). This makes it likely that the short German versions have good construct validity and responsiveness, too. However, although direction and observed dose-response relationships of associations support the validity of the short forms, we found only low to moderate correlations for hypothesized associations between oral health conditions and OHRQoL. This may be because of the fact that OHRQoL is a broad concept reflecting individuals' perceptions about the impact of oral conditions, not simply the presence of the conditions.

Reliability of the 14- and 21-item German short forms was similar to that of the published English-language short forms. The OHIP-G5, the shortest instrument, had an internal consistency of somewhat lower magnitude compared with other OHIP study results (10, 13).

Different methodological approaches to develop a short scale

Interestingly, different methodological approaches for item selection gave rise to similarly good scales from a psychometric point of view. Recently a second English-language OHIP short form was developed where the item selection was based on a different method compared with the original short form (13). The original method to select items was *regression-based* (stepwise and controlled regression). A regression model selected items which predicted the OHIP summary score best according to a criterion. The approach used for the second OHIP short form was a *modified clinical impact method* originally suggested by Juniper et al. (26). The frequency and the severity of items are used to select items with 'impact'. The two short forms developed by these different techniques performed about equally well in our study.

For the five- and 21-item German instruments, other, different item selection methods seemed to perform equally well, too. The method used to create the OHIP-G21 could be characterized as *factor analysis based*. Our method for the OHIP-G5 combined features of the *regression-based* method with the *clinical impact* method. The selection of the dimension indicators does not have to rely solely

on statistical grounds. For instance, we selected the ninth model as the final short form and not the first. 'Difficulty chewing' is a strong indicator for one of the major oral functions – chewing. It is characterized by a considerable prevalence in the general population (23). Secondly, in our opinion 'felt uncomfortable about appearance' seems a better indicator for the dimension *appearance* than 'worried'. Our item selection procedure might be best described as a 'regression-guided item selection'. The statistical method served to discover relationships in the data. By searching through all possible statistical models, best subset regression presented a specified number of models and gave insight into the relationship between variables. Finally, clinical expertise was used to select the most appropriate model among a variety of good models. Prevalence of impacts was considered clinically important similar to the above mentioned clinical impact method. However, what may be considered an advantage – the incorporation of clinical expertise into the selection of items – may also be considered a disadvantage. Item selection affects critically discriminative and evaluative psychometric properties of the instrument. Other authors could have selected a different model if their criteria were different.

The fact that all methods were able to produce instruments with adequate psychometric properties supports the notion that there is a strong underlying construct OHRQoL. We consider the development of short OHIP questionnaires, which describe the construct OHRQoL as a whole as robust against methodological influences.

Limitations

Our study has several limitations. In the validity assessment, we used the same sample from the general population where we validated the long instrument. This may be the most serious limitation of our study. Because the long and the short version are correlated, it was expected *a priori* that the summary scores of the short versions would correlate with conditions of known OHRQoL impact because the summary scores of the long version already did. However, we found identical patterns in a different population of TMD patients. TMD patients may be considered a target population for quality of life assessment because of the considerable psychosocial impact of orofacial pain (27). This sample meets recommendations to use independent samples of subjects from the target population for validation (28, 29). Although the long version of the OHIP was developed in the general population, the concept of OHRQoL should be similar in patient samples with specific oral conditions. The fact that the hypothesized relationships were also found in TMD patients strengthens the validity of OHIP short and long versions.

In our reliability analyses, test-retest data came from the same sample as the long instrument. However, two types of reliability analyses (test-retest and internal consistency) were consistent. Results from a new sample (TMD patients) and the sample from the general population were consistent. According to the Spearman–Brown Formula (30) the OHIP-G with five items should have a lower Cronbach's alpha than the 14-item measures. Those two instruments should have less internal consistency than the 21-item instrument. This ranking was observed. In conclusion, reliability results for the short OHIP followed the hypothesized patterns.

Several methods of item selection to shorten a long instrument are available. Our analyses used items that were not weighted. Disadvantages of weights are that they complicate the derivation of a summary score and they make interpretation more difficult compared with a simple sum of the item responses. Hoped-for improved measurement properties using item weights in cross-sectional or longitudinal studies are not known for the OHIP (19, 31). Using weights might be appropriate for detailed research questions where weighting schemes might improve validity. Based on our findings and literature results, we see no advantages of weighting of OHIP-G items for most purposes.

We presented a smaller set of 21 items which suggested orofacial pain, oral functions, psychosocial impact, and appearance as possible dimensions for OHRQoL. In the current study, we found the OHIP-G21 to be a valid, reliable and responsive instrument to measure the construct OHRQoL. Although promising, our suggested dimensional structure is so far only supported from exploratory factor analytic methods. Therefore, we consider the 21-item version as preliminary. Validation of the dimensions is necessary using conditions with hypothesized OHRQoL dimension-specific impact and confirmatory factor analyses.

Advantage and disadvantage of short versions of the German Oral Health Impact Profile

We expect that short OHIP-Gs will be a useful alternative to the long instrument when time and

resources are limited. However, although brevity is a major advantage, the longer versions might nonetheless be preferred in some contexts because of their greater content validity or when minimizing the degree of random measurement error is especially important. The choice between the three options – the OHIP-G5, the OHIP-G14 [based on the first English-language OHIP-14 (10)] and the OHIP-G21 depends on the purpose of the study.

The OHRQoL can be evaluated on the item level, in dimensions and as construct. The option to characterize OHRQoL on the item level provides useful information. This is severely limited when short versions are used. Saving resources and gaining time should carefully weighed against these disadvantages. Short forms with the possible exception of the OHIP-G21 do not describe dimensions precisely. Therefore, the use of short forms should only be considered if the primary purpose is characterization of the entire construct OHRQoL.

Acknowledgments

The authors are grateful to Jörg Paus and Udo Jellesen (Taylor Nelson Sofres Erforschung der öffentlichen Meinung, Marktforschung, Nachrichten, Informationen, Dienstleistungen) for their help collecting the data. The study was financed by the Institute of German Dentists, Cologne. It was supported by Deutsche Akademie der Naturforscher Leopoldina Grant BMBF-LPD 9901/8-4.

References

1. John MT, Patrick DL, Slade GD. The German version of the Oral Health Impact Profile - translation and psychometric properties. Eur J Oral Sci 2002;110:425–33.
2. Slade GD, Spencer AJ. Development and evaluation of the Oral Health Impact Profile. Community Dent Health 1994;11:3–11.
3. Slade GD, Strauss RP, Atchison KA, Kressin NR, Locker D, Reisine ST. Conference summary: assessing oral health outcomes – measuring health status and quality of life. Community Dent Health 1998;15:3–7.
4. John MT, Hujoel P, Miglioretti DL, Leresche L, Koepsell TD, Micheelis W. Dimensions of oral-health-related quality of life. J Dent Res 2004;83:956–60.
5. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417–32.
6. Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. Qual Life Res 1998;7:323–35.
7. Bullinger M, Anderson R, Cella D, Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Qual Life Res 1993;2:451–9.
8. Locker D, Matear D, Stephens M, Lawrence H, Payne B. Comparison of the GOHAI and OHIP-14 as measures of the oral health-related quality of life of the elderly. Community Dent Oral Epidemiol 2001;29:373–81.
9. Robinson PG, Gibson B, Khan FA, Birnbaum W. A comparison of OHIP 14 and OIDP as interviews and questionnaires. Community Dent Health 2001;18:144–9.
10. Slade GD. Derivation and validation of a short-form oral health impact profile. Community Dent Oral Epidemiol 1997;25:284–90.
11. Bowling A. Measuring health: a review of disease-specific quality of life measurement scales. 2nd edn. Buckingham: Open University Press; 2001. p. v–vii.
12. Locker D. Applications of self-reported assessments of oral health outcomes. J Dent Educ 1996;60:494–500.
13. Locker D, Allen PF. Developing short-form measures of oral health-related quality of life. J Public Health Dent 2002;61:13–20.
14. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord 1992;6:301–55.
15. Lipton JA, Ship JA, Larach-Robinson D. Estimated prevalence and distribution of reported orofacial pain in the United States. J Am Dent Assoc 1993;124:115–21.
16. Murray H, Locker D, Mock D, Tenenbaum HC. Pain and the quality of life in patients referred to a craniofacial pain unit. J Orofac Pain 1996;10:316–23.
17. Frahn G, John M. Schmerzen im orofazialen System – eine kontrollierte Studie mit Stabilisierungsschiene und Ultraschall. Dtsch Zahnärztl Z 1996;8:478–81.
18. Skevington SM. Investigating the relationship between pain and discomfort and quality of life, using the WHOQOL. Pain 1998;76:395–406.
19. Allen PF, McMillan AS, Locker D. An assessment of sensitivity to change of the Oral Health Impact Profile in a clinical trial. Community Dent Oral Epidemiol 2001;29:175–82.
20. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
21. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
22. Cronbach LJ. Coefficient alpha and the internal reliability of tests. Psychometrika 1951;16:297–334.
23. John MT, LeResche L, Koepsell TD, Hujoel PP, Miglioretti DL, Micheelis W. Oral health-related quality of life in Germany. Eur J Oral Sci 2003;111:483–91.
24. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd edn. New York: John Wiley & Sons; 2003. p. 598–628.
25. Allison P, Locker D, Jokovic A, Slade G. A cross-cultural study of oral health values. J Dent Res 1999;78:643–9.
26. Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of

life questionnaire construction. J Clin Epidemiol 1997;50:233–8.
27. Dworkin SF. Behavioral, emotional, and social aspects of orofacial pain. In: Stohler CS, Carlson DS, editors. Biological and psychological aspects of orofacial pain. Ann Arbor: Center for Human Growth & Development, The University of Michigan; 1994. p. 93–112.
28. Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. Stat Med 1995;14:331–45.
29. Coste J, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. J Clin Epidemiol 1997;50:247–52.
30. Streiner DL, Norman GR. Health measurement scales. 2nd edn. Oxford: Oxford University Press; 1995. p. 104–27.
31. Allen PF, Locker D. Do item weights matter? An assessment using the oral health impact profile. Community Dent Health 1997;14:133–8.