Developing Short-form Measures of Oral Health-related Quality of Life

David Locker, PhD; P. Finbarr Allen, PhD

Abstract

Objectives: *Using the item-impact method, we developed an alternative short-form Oral Health Impact Profile (OHIP) that has good psychometric properties and minimal floor effects.* ***Methods:*** *OHIP data were collected from a sample of older Canadians at two points in time. Data from the first administration were used to develop a 14-item short-form measure; data from the second compare the latter's psychometric properties with those of the original short form developed by Slade (1997), who used a controlled regression procedure.* ***Results:*** *The short form based on the item-impact method had only two items in common with the short form derived from the regression approach and contained more high-prevalence items. The regression short form was subject to marked floor effects, while the impact short form had floor effects comparable to those of the full 49-item OHIP. The former discriminated between dentate and edentulous subjects, while the latter did not. Both discriminated between dentate subjects who did and did not wear dentures, those with and without dry mouth, and those with and without chewing problems. Both were also significantly associated with self-ratings of oral health, satisfaction with oral health, and self-perceived need for dental treatment. The strength of the associations was somewhat stronger with the regression short form, indicating that it performed better as a discriminatory instrument. However, because of its floor effects, it was markedly less sensitive to change than the impact short form. There was an indication that item-impact methods of shortening oral health-related quality of life measures produced more stable results across samples than the statistical approach.* ***Conclusions:*** *Because the content validity of short-form measures is always compromised, different short forms are required for different purposes and different patient populations. The regression short form developed by Slade (1997) is likely to be better when the aim is to discriminate, while the impact short form developed here may be preferable when the aim is to describe the oral health-related quality of life of populations or to detect change. [J Public Health Dent 2002;62(1):13-20]*

Key Words: oral health-related quality of life, outcome measures, reliability, validity, responsiveness, psychometric methods.

Over the last 10 years, a number of measures have been developed to assess oral health-related quality of life (OHRQL) (1). These are similar in most respects to measures developed in medicine in that they were designed to document the functional, psychological, and social impact of diseases and disorders affecting the oral cavity and related structures. The majority takes the form of composite measurement scales (2) and consists of selected items that address conceptually distinct dimensions of health and well-being. These measures have a number of potential applications in research, clinical, and public health practice (3). In medicine they are increasingly used to evaluate the health of populations, to compare the outcomes of different therapeutic interventions, and to make decisions about the care of individual patients (2). In dentistry, measures of oral health-related quality of life have been used in oral health surveys of adult, older adult, and elderly populations (4-6), and as outcome measures in clinical trials of implant therapies (7,8) and evaluations of dental care programs for special care populations (9). Their use in clinical practice and clinical decision making has yet to be reported.

Although many measures of health and oral health-related quality of life have good psychometric properties and have proved to be reliable, valid, and responsive to clinically meaningful change, their use in some settings may be limited by their length and the complexities involved in completing and scoring the instruments. Measures that take a long time to complete or score may not be feasible in clinical settings because of the burden placed on patient and/or clinician. They may be inappropriate for use in national population surveys because of the increased costs of data collection or in smaller scale surveys where numerous other general health and psychological measures are being used. Respondent burden may be an issue when the participants in a study are frail, ill, or severely compromised and unable to cooperate for any length of time. Long scales also are more likely to be subject to item nonresponse, giving rise to problems of how to manage missing data. For these reasons, many investigators have attempted to increase the cost efficiency of assessing health-related quality of life by the development of short-form instruments (2).

One of the most sophisticated of the OHRQL measures developed to date is the Oral Health Impact Profile (OHIP) (10). This is a generic 49-item measure consisting of seven subscales. The seven dimensions addressed by the measure are: functional limitations, physical pain, psychological discomfort, physical disability, psychological disability, social disability, and handicap. The OHIP was based on a coherent conceptual framework and

Send correspondence and reprint requests to Dr. Locker, Faculty of Dentistry, University of Toronto, 124 Edward Street, Toronto, Ontario M5G 1G6, Canada. E-mail: david.locker@utoronto.ca. Dr. Allen is with the Department of Restorative Dentistry, University Dental School and Hospital, National University of Ireland, Cork. Manuscript received: 2/2/01; returned to authors for revision: 4/18/01; accepted for publication: 7/6/01.

items were derived from qualitative interviews with patients suffering from a wide variety of oral disorders. It is one of the few measures in which each item has a weight that indicates the severity of the problem described by the item. Consequently, scores can be derived in a number of ways (11), with high scores indicating poor OHRQL. Slade (12) has also developed a short form of the OHIP (OHIP-14), which consists of 14 items. This was developed using least squares regression, with the total OHIP score as the dependent variable and a controlled stepwise procedure in which the two items from each subscale making the largest contribution to the R^2 were selected. This method of selecting items proved to be more satisfactory than internal reliability analysis or factor analysis.

Using data from an Australian study of older adults, Slade (12) demonstrated that the OHIP-14 was as reliable and valid as the OHIP-49. Similar conclusions regarding reliability and validity were reached when the short form was used in a study of an elderly Canadian population resident in a long-term care facility (13). The OHIP-14 had an internal consistency reliability of 0.87 and showed good concurrent and construct validity. However, when used with this very old and medically compromised population, 31 percent of whom were edentulous, the measure had significant "floor" effects (14). That is, 30.3 percent to 45.8 percent of subjects overall had a score of 0, depending upon the method of calculating scores, as did 17.8 to 35.6 percent of subjects rating their oral health as only fair or poor. This means that the measure would be unable to detect improvements in OHRQL in a large minority of this population following an intervention of known efficacy.

This floor phenomenon may have arisen because the development of a short-form measure must of necessity compromise content validity. The regression approach used to derive the OHIP-14 excluded items describing common problems such as "difficulty chewing," "food catching," and "sensitive teeth," while some severe and infrequent items, such as "difficulty doing usual jobs" and being "totally unable to function," were selected. In fact, of the 14 items selected, only five were among the two most commonly reported within their respective subscales. As Slade (12) indicated, the inclusion of low-frequency, relatively severe items will maximize the ability of the measure to discriminate between groups with low and high levels of disease. However, the question then arises as to whether a different approach to developing short-form measures would give rise to a different subset of items and whether this alternative short form would be less subject to floor effects, while maintaining good reliability and validity. The reduction of floor effects would facilitate the use of the short form in clinical practice, clinical trials, and evaluative studies by enhancing its ability to detect clinically meaningful change.

Coste et al. (2) and Juniper et al. (15) discuss three philosophically different ways of shortening composite measurement scales. These are statistical methods, the expert-based approach, and the item-impact method. The first tends to be the most common method used, with factor analysis and regression methods predominating. Here, items for the short form are selected based on their relationships with other items or their ability to predict overall scale scores. As the name suggests, the expert-based approach uses the judgments of experts who are knowledgeable about a disease and its consequences for HRQL. Experts such as health care providers often are used to generate the initial item pool out of which an original measure is developed (16). This approach can be limited by the fact that expert opinions are not necessarily objective or sufficiently representative.

The item-impact method selects items that are the most important to patients. This method usually is used at an earlier stage of the process of developing a health-related quality-of-life questionnaire to select items from an initial item pool for a final questionnaire, which is then tested for reliability and validity (17). Samples of patients are given the initial item pool and asked to indicate which of the items describe problems they have experienced in the previous year. For each item identified, they are asked to rate its importance on a five-point scale ranging from "not important" to "very important." Item-impact scores are developed by multiplying the proportion of patients experiencing the item by its mean importance rating. Items then are ranked according to these scores and the highest scoring items selected. Clearly, this method can be adapted readily to developing short-form instruments from longer instruments. Here, the long form of the measure is treated as if it were an initial item pool. Juniper et al. (15) compared the item-impact method and factor analysis in reducing a pool of 152 items describing problems experienced by individuals with asthma. The former resulted in an instrument with 32 items and the latter an instrument with 36 items. Although 20 of the items were common, the two approaches produced quite different measures.

Because of its intuitive appeal, we used the item-impact method with data from a population-based study of older Canadians to develop an alternative short form of the OHIP. The goal was to derive a 14-item measure with a simple scoring system that could be used readily in clinical contexts to measure changes in patient well-being. This paper describes the development and evaluation of this alternative short form and compares its measurement properties with those of the original short form developed by Slade (12).

Methods

Data Sources. The data used to develop and evaluate the short-form OHIP comprise part of the Ontario Study of the Oral Health of Older Adults. This is an observational cohort study of a random sample of community-dwelling persons aged 50 years and older at baseline. Data were collected at baseline and at one-, three-, and seven-year follow-ups using personal interviews and clinical examinations. Details of this study have been reported previously (5,18,19). Self-completed versions of the 49-item OHIP were used at the follow-ups at year 1 and year 3. At the one-year follow-up, data were obtained from 699 subjects; at three years, data were obtained from 541 subjects. For each of the 49 OHIP items, respondents were asked to indicate how often over the previous year they had experienced the problem described by the item using the following Likert frequency response scale: never=0, hardly ever=1, occasionally=2, fairly often=3, very often=4. A "don't know" option also was included. Data from the one-year follow-up were used to calculate item-

impact scores and to select items for the short form, while data from the three-year follow-up were used to assess reliability and validity and other properties such as floor effects. At each data collection point, information on the personal and oral health characteristics of subjects were obtained from personal interviews and oral examinations.

Calculation of Item-impact scores and Selection of Items. Using data collected at the year 1 follow-up, an impact score was calculated for each OHIP item after deleting data for subjects with missing or "don't know" responses. Since an importance rating for each item was not obtained, we used the following method to obtain item-impact scores. For each item the proportion of subjects with response codes 1 (hardly ever) through 4 (very often) was calculated along with the mean frequency rating for subjects with these codes. This mean frequency rating was obtained by summing the response codes for subjects responding "hardly ever" to "very often" and dividing this sum by the number of subjects with those responses. An item-impact score was obtained by multiplying these two values and then multiplying by the item weight. Canadian weights for the OHIP have been developed by Allison et al. (20). Scores were ranked within OHIP subscales and the two top scoring items in each subscale selected for the modified short form. The content of this alternative short form was then compared with that of the original OHIP-14. To facilitate the presentation of the results of this comparison, the alternative was designated the impact short form while the original OHIP-14 was designated the regression short form. These names were selected to reflect the way in which the two versions were developed.

Reliability, Validity, and Floor Effects. Assessments of the psychometric properties of the two short-form OHIPs were undertaken using data collected during the three-year follow-up of the Ontario Study of Older Adults. Two methods were used to derive scores. Additive scores (ADD) were obtained by summing the response codes for the 14 items comprising each version. Simple count (SC) scores were created by counting the number of items with responses "occasionally," "fairly often," or "very often." ADD scores could range from 0 to 64 and SC scores from 0 to 14, with higher scores indicating worse oral health-related quality of life. ADD and SC scores were also calculated for the full 49-item OHIP.

Floor effects for both measures and both scoring methods were assessed by calculating the percentage of subjects with 0 scores. These calculations were made for all subjects and for those rating their oral health as only fair or poor. Internal consistency reliability of the two versions was assessed using Cronbach's alpha. Content validity was assessed by means of the correlations between short-form scores and scores based on the total set of 49 OHIP items. Least squares regression analyses were undertaken using the OHIP-49 score as the dependent variable to determine the contribution of each short form to total R^2. Validity was assessed using tests of discriminative and concurrent validity. This involved a comparison of the strength of the associations between scores derived from the measures and a number of variables designed to indicate the oral health characteristics of this population. The analytic approach used in the paper is similar to that previously used (11,13) in comparisons of the performance of different measures of oral health-related quality of life.

The discriminative validity of two versions was assessed by means of their associations with dental status, partial denture wearing (dentate only), dry mouth, and a problem chewing. Concurrent validity was assessed by means of associations with self-rated oral health, dissatisfaction with oral health status, and self-perceived need for dental treatment. Because scores from both short forms were highly skewed and could not be normalized using log transformations, Mann-Whitney tests were used to assess associations between scores and these other variables. As this test is based on ranks, the differences in mean ranks between categories of the independent variables were used to compare the ability of the measures to distinguish between groups. In addition, scores were dichotomized using median splits, cross-tabulations performed, and odds ratios calculated. These odds ratios provided a more readily interpretable measure of the strength of the associations between dependent and independent variables.

The ability of the two measures to detect change was assessed by means of change scores and their associations with global transition judgments of change (21) collected at the three-year phase. Change scores were obtained by subtracting scores obtained at the three-year follow-up from scores obtained at the one-year follow-up. Consequently, positive scores indicated an improvement in oral health-related quality of life. Mean change scores and effect sizes were calculated for subjects who reported that their oral health had improved. An effect size is a distribution-based measure of the amount of change detected and is calculated by dividing the change score by the standard deviation of the initial score. Effect sizes of 0.2 are considered to be small, 0.6 as moderate, and 0.8 as large (22).

Finally, to determine the stability of the two item-reduction methods, the controlled regression analysis was repeated using the one-year follow-up OHIP data. The item-impact method then was applied to the Australian older adult population used in developing the original OHIP-14, using prevalence and mean frequency data reported in Slade (12) and the Australian item weights reported by Slade and Spencer (10). The item content of these versions was then compared.

Results

Item-impact scores derived from year-one OHIP data ranged from 16.2 to 235.0, reflecting the wide variation in prevalence and severity of items. Of the 14 items with the highest impact scores three came from the functional limitations subscale, four from physical pain, five from psychological discomfort, and one each from physical disability and psychological disability. Consequently, to ensure representation from all seven subscales and maximize content validity, the two top scoring items from each were selected for the impact short-form OHIP.

Table 1 lists the 14 items in the regression and impact short-form OHIPs. The two short forms had only two items in common. Evaluation of the two short forms was conducted using 435 subjects with no missing data on the items comprising both measures. The proportions of subjects at the three-year follow-up with re-

sponses "occasionally," "fairly often," and "very often" confirmed that the regression short form contained more low frequency items than the impact short form (Table 1). For the former, prevalences ranged from 1.3 percent to 27.9 percent. Six had prevalences of less than 10 percent and only two had prevalences of 20 percent or greater. For the latter, prevalences ranged from 4.7 percent to 78.3 percent, with five being below 10 percent and six 20 percent or greater.

Scores on the impact short form developed here were significantly higher than on the regression short form developed by Slade (1997). Using additive scores, means were 8.1 and 4.2, respectively ($P<.001$; paired t-test) and medians were 6.0 and 2.0, respectively ($P<.0001$; Wilcoxon signed rank test). Using simple count scores, means were 2.6 and 1.1, respectively ($P<.001$; paired t-test), while medians were 2 and 0 ($P<.0001$; Wilcoxon signed rank test).

Data pertaining to floor effects are shown in Table 2. For the full 49-item version of the OHIP, 2.0 percent had an additive score of 0 and 13.8 percent had a simple count score of 0. Using ADD scores, 33.1 percent of all subjects had a score of 0 on the regression short form, compared with 2.5 percent for the impact short form. Among those with a score of 0 on the regression short form, 92.4 percent had a score of 1 or more on the impact short form (mean=3.5, median=4.0) and 95.5 percent had an OHIP-49 score of 1 or more (mean=5.3, median=5.0). Using SC scores, 58.9 percent had a regression short form score of 0 and 13.6 percent had an impact short form score of 0. Again, four-fifths of those with a score of 0 on the regression version had a score of one or more on the impact version (77.3%; mean=1.3, median=1.0) and OHIP-49 (80.0%; mean=1.8, median=1.0). The regression short form also showed marked floor effects for those subjects rating their oral health as fair or poor. Using ADD scores, 14.3 percent had a score of 0; using SC scores a score of 0 was observed for 28.6 percent. Figures for the impact short form were 3.6 percent and 3.6 percent.

Internal consistency reliabilities were 0.91 for the regression short form and 0.85 for the impact short form. Spearman's rank correlations between short-form scores and the scores derived from the 49-item OHIP were similar for both versions and scoring methods, ranging from 0.90 to 0.94 ($P<.001$). In least squares regression analyses using additive scores, the regression short form accounted for 94 percent of the variance in OHIP-49 scores, while the impact short form accounted for 91 percent. In both cases, one of the 14 items did not enter the regression model.

Tables 3 and 4 provide data on the discriminant and concurrent validity of the two short-form measures. Scores from the regression short form discriminated between dentate and edentulous subjects, dentate subjects who did and did not wear partial dentures, subjects with and without dry mouth, and subjects with and without a problem chewing. Scores from the impact short form did not discriminate between dentate and edentulous subjects, but did discriminate between subgroups defined by denture wearing, dry mouth, and a chewing problem. Differences in mean ranks obtained from Mann-Whitney tests indicated that where both versions discriminated between groups, the re-

TABLE 1
Item Content of Short-form OHIPs and Percent Reporting Each Item Occasionally, Fairly Often, or Very Often

Subscale	Regression Short Form	%*	Impact Short Form	%*
Functional limitation	Trouble pronouncing words	11.4	Difficulty chewing	25.7
	Sense of taste worse	10.8	Food catching	78.3
Physical pain	Painful aching in mouth	18.8	Sensitive teeth	31.0
	Uncomfortable to eat	27.9	Sore spots	34.8
Psychological discomfort	Self-conscious	20.3	Worried	27.0
	Tense	13.6	Miserable	14.2
Physical disability	Diet unsatisfactory	5.6	Speech unclear	9.8
	Had to interrupt meals	10.3	Avoid eating some foods	20.5
Psychological disability	Difficult to relax	7.6	Been upset	14.4
	Been embarrassed	13.2	Been depressed	8.0
Social handicap	**Irritable with others**	4.7	Less tolerant of others	2.3
	Difficulty doing jobs	1.3	**Irritable with others**	4.7
Handicap	**Life less satisfying**	9.5	Financial disadvantage	13.4
	Totally unable to function	1.9	**Life less satisfying**	9.5

*Percent reporting each item occasionally, fairly often, very often.

TABLE 2
Floor Effects: Percent of Subjects with 0 Scores According to Version and Scoring Method

Measure	Additive Method (%)	Simple Count Method (%)
All subjects		
OHIP-49	2.0	13.8
Regression short form	33.1	58.9
Impact short form	2.5	13.6
Subjects rating oral health as fair or poor		
OHIP-49	1.4	1.4
Regression short form	14.3	28.6
Impact short form	3.6	3.6

TABLE 3
Discriminant Validity: Median Values for Each Category of Grouping Variable, Differences in Mean Ranks, and Odds Ratios

	Additive Scores		Simple Count Scores	
	Regression Short Form	Impact Short Form	Regression Short Form	Impact Short Form
Dental status				
Edentulous	3.0	6.0	1.0	2.0
Dentate	2.0	6.0	0.0	2.0
P-value*	<.01	—	<.05	—
Diff. in mean ranks	48	–2	39	–3
Odds ratio	1.7	0.9†	1.7†	1.1†
Partial denture‡				
Yes	3.0	7.0	0.0	2.0
No	2.0	6.0	0.0	2.0
P-value	<.001	<.01	<.01	<.05
Diff. in mean ranks	38	31	32	31
Odds ratio	2.0	1.8	1.9	1.6
Dry mouth				
Yes	3.0	7.0	0.0	2.0
No	2.0	6.0	0.0	2.0
P-value	<.001	<.01	<.01	<.05
Diff. in mean ranks	41	37	32	31
Odds ratio	1.9	1.6	1.8	1.6
Chewing problem				
No	6.0	10.0	1.5	4.0
Yes	2.0	6.0	0.0	2.0
P-value	<.0001	<.001	<.0001	<.0001
Diff. in mean ranks	112	85	94	75
Odds ratio	4.2	2.7	4.3	3.4

*From Mann-Whitney tests.
†95% confidence interval includes 1; 95% confidence intervals for all other odds ratios exclude 1.
‡Dentate only.

gression short form was marginally better. Odds ratios based on median splits also indicated that the associations between scores and the four independent variables were stronger for the regression than the impact version OHIP. This indicates that the former version was better at discriminating between groups than the latter.

Scores derived from both short forms were significantly associated with self-rated oral health, dissatisfaction with oral health, and self-perceived need for dental treatment. Again, most of the associations were marginally stronger with scores derived from the regression short form than those from the impact short form. This is to be expected; those with higher scores on the original short form are more likely to experience one or more of the low prevalence higher severity impacts and would, therefore, be more likely to rate their oral health as poor and/or be dissatisfied with their oral health status.

Change scores for both versions and both scoring methods were significantly associated with subjects' global transition judgments ($P<.001$ for all analyses). However, mean change scores and effect sizes for subjects reporting that their oral health had improved were higher for the impact short form developed here than the regression short form (Table 5). This was particularly the case when SC scores were used. The effect size was 0.48 for the regression short form (small to moderate) and 1.1 (strong) for the impact short form.

Finally, when the controlled regression method of selecting items was used with these Canadian data, the resulting measure contained only seven items from the original short form and five of these came from the functional limitations and physical pain subscales. When the item-impact method was used with Australian data, 11 of the 14 items were the same as when this method was used with Canadian data. This suggests that the item-impact approach to item reduction produces a more stable result across samples than the statistical approach based on regression analysis.

Discussion

The main aim of the study reported here was to determine if a short-form OHIP developed using a modified form of the item-impact method resulted in a measure that maintained the psychometric properties of the original short form described by Slade (12), while minimizing floor effects. A secondary aim was to compare the content of measures produced by different approaches to the development of short-form instruments.

The item-impact approach to selecting items for short-form questionnaires involves the calculation of a score for each item. That score is obtained by multiplying the prevalence of the item in the population of interest by some measure of its importance to that population. The preferred method of assessing importance is to ask members of the population of interest to rate the importance of the items that apply to them. Since we did not collect importance ratings, we modified that approach and assessed importance using data on the frequency with which an item was experienced and its severity as indicated by the item weight. This seems justified because frequency and severity are likely to be taken into account when individuals rate the importance of problems they experience as a result of oral disorders. However, it is possible that direct ratings of importance would have resulted in different item-impact scores and a different set of questions being selected for the impact short form.

Another point that needs to be made is that item-reduction procedures always involve a combination of statistical considerations and subjective judgment (15). In developing the original short form, Slade (12) eliminated items with high nonresponse rates and items specific to denture wearing. The particular regression approach used also ensured that only two items per

TABLE 4
Concurrent Validity: Median Values for Each Category of Grouping Variable, Differences in Mean Ranks, and Odds Ratios

	Additive Scores		Simple Count Scores	
	Regression Short Form	Impact Short Form	Regression Short Form	Impact Short Form
Self/rated oral health				
Fair/poor	6.0	11.0	2.0	4.0
Excellent	2.0	6.0	0.0	2.0
P-value	<.0001	<.0001	<.0001	<.0001
Diff. in mean ranks	108	107	102	98
Odds ratio	5.6	5.0	4.9	4.5
Dissatisfied with oral health				
Yes	8.0	12.0	3.0	4.0
No	1.5	6.0	0.0	2.0
P-value	<.0001	<.0001	<.0001	<.0001
Diff. in mean ranks	141	118	133	104
Odds ratio	8.7	6.9	8.5	4.7
Self-perceived need for treatment				
Yes	4.0	8.0	1.0	3.0
No	2.0	6.0	0.0	2.0
P-value	<.0001	<.0001	<.0001	<.0001
Diff. in mean ranks	68	74	68	72
Odds ratio	3.1	2.8	2.9	2.7

*From Mann-Whitney tests.
95% confidence intervals around all odds ratios exclude 1.

TABLE 5
Mean Change Scores and Effect Sizes: Subjects Reporting Improved Oral Health

	Additive Scores		Simple Count Scores	
	Regression Short Form	Impact Short Form	Regression Short Form	Impact Short Form
Mean change score	4.7	7.8	1.2	3.0
Effect size	0.62	1.0	0.48	1.0

OHIP subscale were selected. In an attempt to maximize content validity we also selected two items per subscale, rather than taking the 14 items with the highest impact scores.

The item-impact method used here produced a very different short-form instrument than that which emerged out of the regression approach. Our impact short form had only two items in common with the regression short form and contained more items whose prevalence exceeded 20 percent. Consequently, scores on the impact version were significantly higher than scores on the regression version (medians of 6.0 and 2.0, respectively, when using the additive scoring method), indicating that it was identifying more oral health impact. If the aim had been to maximize scores, then it would have been better to use the 14 items with the highest impact scores. The median additive score for this combination of items was 8.0, significantly higher than scores from both of the short forms assessed here.

Both short forms showed excellent internal consistency reliability when used with this population. There was also a high correlation between scores from both short forms and scores from the full 49-item version of the OHIP. Moreover, the items comprising the two short forms explained 94 percent and 91 percent, respectively, of the variance in total OHIP scores. However, undue emphasis should not be placed on these R^2 values, since randomly selected blocks of 14 items had R^2 values ranging from 0.86 to 0.94. This is because all 49 OHIP items showed significant and moderate to strong correlations with total scale scores. This suggests that any subset of 14 items will probably have reasonable psychometric properties, however selected.

Both short forms performed well when tested for discriminant and concurrent validity. The regression short form was better at discriminating between groups and distinguished between subpopulations based on dental status, denture wearing (among the dentate), dry mouth, and oral dysfunction. The impact short form did not discriminate between dentate and edentulous subjects. Although scores were associated with denture wearing, dry mouth, and oral dysfunction, relationships were less strong than those achieved with the regression short form. Most of the relationships between the impact short form we developed and global indicators of oral health-related quality of life—such as self-rated oral health, dissatisfaction with oral health status, and self-perceived need—were also less strong than those observed with the regression short form. However, when used with this general population of older adults, the version developed using the impact method was far less subject to floor effects and showed greater sensitivity to change.

The superior ability of the regression short form to discriminate between groups stems from the fact that the majority of its items were reported by fewer than 20 percent of the sample we studied. Juniper et al. (23) suggest that high prevalence items should not be included in a discriminatory measure, since they compromise its ability to distinguish between groups with severe and less severe disease. Hyland et al. (24) also stipulate that items endorsed by 70 percent or more of respondents will be poor discriminators and should be discarded. The analysis presented here indicates that this is also the case with measures of oral health-related quality of life. How-

ever, Guyatt et al. (23) suggest that a basic principle when constructing measures designed to detect change is that they should be based on what patients feel is most important. This is why they use the item-impact method in developing measures for use in clinical trials, because this specifically identifies those items that are most frequently experienced and impact on patients to the greatest degree (17). Again, the analysis presented here suggests that including high-frequency items facilitates the measurement of change in oral health. Clearly, an instrument that successfully discriminates between groups may not be optimal at detecting change and may not be appropriate as an outcome measure in studies of health care interventions.

This suggests that different short forms may be needed according to the purpose for which the measure is being used. There are, in essence, three types of measure, each of which performs different functions. Descriptive measures are used in surveys to document population oral health-related quality of life. The aim here should be to maximize scores. Consequently, a measure consisting of low-prevalence items will fail to document the full extent to which oral conditions impact on populations. As the analysis presented here demonstrated, the majority of subjects with 0 scores on the regression short form did in fact experience some impact from oral disorders as evidenced by their scores on the impact short form and the total 49-item OHIP. The best short form for population surveys may then be a measure consisting of items with the highest impact scores.

Discriminative measures are used in clinical contexts to differentiate between groups with different conditions or conditions of different levels of severity. As noted above, measures that consist of items affecting most patients will fail to distinguish between those who are and are not severely compromised. Evaluative measures are used to assess the extent of within-subject change that occurs as a result of health care interventions. Since health care interventions should be targeted toward what patients feel is important, it is essential that these are measured precisely (25). Consequently, evaluative instruments need to contain an adequate representation of high-frequency items. Juniper at al (26) developed an Asthma Quality of Life Questionnaire consisting of 32 items. All had prevalences greater than 40 percent and 12 had prevalences greater than 70 percent. Low-frequency items are likely to describe severe impacts that are less amenable to change; including too many of these items is likely to compromise an instrument's responsiveness (15,27).

Evaluative measures should also reflect the specific goals of an intervention (17). For example, prosthodontic interventions may be designed to improve chewing capacity and improve eating and diet, so that items relevant to these goals should be included. Items relating to pain would be of less relevance in this context. A study of interventions to reduce the psychosocial consequences of chronic facial pain should contain items on pain and the psychological outcomes of that pain. Questions concerning appearance and embarrassment will be of less relevance in this context.

One important limitation of the study reported here, and that of Slade (12), is that these short form OHIPs have been developed and evaluated using general population samples with relatively common oral disorders rather than samples of patients with specific and/or severe clinical conditions. Consequently, conclusions regarding their reliability and validity, floor effects, and sensitivity to change may not apply to patients with such disorders. For example, when used with facial pain patients, the original OHIP-14 may not show the marked floor effects evident in this study, since low-prevalence items may become high-prevalence items. In a study of chronic facial pain patients that used the OHIP, 24.6 percent responded positively to the item "Totally unable to function" (28). If the findings can be applied to patient populations, they may be limited to patients attending general dental practice for the treatment of routine dental and oral conditions who approximate general population samples. In this situation, the impact short form may be preferable to the original version. Consequently, these alternate short forms need to be tested in clinical contexts on samples of patients with a variety of specific disorders. In particular, their relative sensitivity to change should be assessed in clinical studies of known efficacy. Ideally, however, disease-specific short forms should be developed using data from patients with the disease of interest rather than using data derived from general populations.

It is also the case that the impact short form developed here should be tested on other general population samples. As a general principle, short forms should be developed using one sample and the psychometric properties of the measure tested on a new and independent sample (2). While our measure was developed and tested using different data sets collected at different time periods, the sample remained the same. Consequently, while the analyses described above have confirmed the reliability and validity of Slade's original short form based on regression, and its performance as a discriminatory instrument, the properties of the short form derived from the impact method need to be explored further.

One solution to the problems discussed in this paper is to use the full 49-item OHIP. Where this is not possible, investigators need to think carefully about the aims and objectives of their study and select a subset of items accordingly. In this respect, the 49-item OHIP provides a valuable resource on which investigators can draw. Moreover, Juniper et al (15) are of the opinion that each dimension on a health status questionnaire needs to be represented by three or four items. This decreases the variability in responses found even in patients whose condition is stable and minimizes the effects of idiosyncratic responses to individual items. In addition, increasing the number of questions per subscale is one way of reducing floor effects in a short-form questionnaire (14). Consequently, investigators should also consider whether or not a 21- or 28-item version of the OHIP may not be preferable or whether some subscales might need more than two items. This would increase the content validity of the instrument by including items referring to toothache and embarrassment, which had the third highest impact scores within their respective subscales.

A final consideration is whether the statistical or item-impact approach is best for item reduction or the development of short-form instruments. Fayers and Hand (29) take the view that most health status measures, because

of their internal causal structure, violate the assumptions of factor analysis and Coste et al. (2) are critical of the use of statistical approaches alone and suggest a combination of statistical considerations and expert opinion. The analyses presented here suggest that item-impact methods may produce a more stable result across samples than statistical approaches. However, Juniper et al (15) are of the opinion that the answer is largely philosophical and depends on the extent to which an investigator believes that there must be mathematical links between the items on a questionnaire. Probably the method of developing a short-form instrument is not as important as its content. In the final analysis, the items in a short-form questionnaire and its measurement properties need to be appropriate to its purpose, the population to which it is applied, and the context in which it is being used.

References

1. Slade GD, ed. Measuring oral health and quality of life. Chapel Hill: University of North Carolina, Dental Ecology, 1997.
2. Coste J, Guillemin F, Pouchot J, Femanian J. Methodological approaches to shortening composite measurement scales. J Clin Epidemiol 1997;50:247-52.
3. Locker D. Applications of self-reported assessments of oral health outcomes. J Dent Educ 1996;60:494-500.
4. Atchison K, Dolan T. Development of the Geriatric Oral Health Assessment Index. J Dent Educ 1990;54:680-7.
5. Locker D. The burden of oral disorders in an older adult population. Community Dent Health 1992;9:109-24.
6. Slade G, Spencer J, Locker D, Hunt R, Strauss R. Variations in the social impact of oral conditions among older adults in South Australia, Ontario, and North Carolina. J Dent Res 1996;75:1439-50.
7. Awad M, Locker D, Korner-Bitensky N, Feine J. Measuring the effect of implant rehabilitation on health-related quality of life in a randomized clinical trial. J Dent Res 2000;79:1659-63.
8. Allen PF, McMillan AS. The impact of tooth loss in a denture-wearing population: an assessment using the Oral Health Impact Profile. Community Dent Health 1999;16:176-80.
9. Dolan T. The sensitivity of the Geriatric Oral Health Assessment Index to dental care. J Dent Educ 1997;61:37-46.
10. Slade G, Spencer A. Development and evaluation of the Oral Health Impact Profile. Community Dent Health 1994;11:3-11.
11. Allen PF, Locker D. Do weights matter? An assessment using the Oral Health Impact Profile. Community Dent Health 1997;14:133-8.
12. Slade G. Derivation and validation of a short-form oral health impact profile. Community Dent Oral Epidemiol 1997; 25:284-90.
13. Locker D, Matear D, Stephens M, Lawrence H, Payne B. Comparison of the GOHAI and OHIP-14 as measures of the OHRQoL of the elderly. Community Dent Oral Epidemiol 2001;29:373-81.
14. Bindman A, Keane D, Lurie N. Measuring health changes among severely ill patients: the floor phenomenon. Med Care 1990;28:1142-51.
15. Juniper E, Guyatt G, Streiner D, King D. Clinical impact versus factor analysis for quality of life questionnaire construction. J Clin Epidemiol 1997;50:233-8.
16. Streiner D, Norman G. Health measurement scales: a practical guide to their development and use. Oxford: Oxford Medical Publications, 1989.
17. Guyatt G, Bombarider C, Tugwell P. Measuring disease specific quality of life in clinical trials. CMAJ 1986;134:889-95.
18. Locker D. Effects of nonresponse on estimates derived from an oral health survey of older adults. Community Dent Oral Epidemiol 1993;23:108-13.
19. Payne B, Ford J, Locker D. Loss to follow-up in a longitudinal survey of older adults. Community Dent Oral Epidemiol 1995;23:297-302.
20. Allison P, Locker D, Jokovic A, Slade D. A cross-cultural study of oral health values. J Dent Res 1999;78:643-9.
21. Locker D. Issues in measuring change in self-perceived oral health status. Community Dent Oral Epidemiol 1998;26:41-7.
22. Cohen J. Statistical power for the behavioral sciences. New York: Academic Press, 1977.
23. Juniper E, Guyatt G, Jaeschke R. How to develop and validate a new health-related quality of life instrument. In: Spiker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven Publishers, 1996.
24. Hyland M, Finnis S, Irvine S. A scale for assessing quality of life in adult asthma sufferers. J Psychosomatic Res 1991; 35:99-110.
25. Mitchell A, Guyatt G, Singer J, et al. Quality of life in patients with inflammatory bowel disease. J Clin Gastroenterol 1988; 10:306-10.
26. Juniper EF, Guyatt GH, Epstein RS, Ferrie PJ, Jaeschke R, Hiller TK. Evaluation of impairment-related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax 1992;47: 76-83.
27. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27-36.
28. Murray H, Locker D, Mock D, Tenebaum HC. Pain and the quality of life in patients referred to a craniofacial pain unit. J Orofac Pain 1996;10:316-23.
29. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Qual Life Res 1997;6:139-50.