*Copyright © Blackwell Munksgaard 2004*

COMMUNITY DENTISTRY AND ORAL EPIDEMIOLOGY

# Assessing the responsiveness of measures of oral health-related quality of life

**David Locker, Aleksandra Jokovic and Martha Clarke**

Community Dental Health Services Research Unit, Faculty of Dentistry, University of Toronto, Toronto, Canada

Locker D, Jokovic A, Clarke M. Assessing the responsiveness of measures of oral health-related quality of life. Community Dent Oral Epidemiol 2004; 32: 10–8.
© Blackwell Munksgaard, 2004

Abstract – *Objectives:* This paper illustrates ways of assessing the responsiveness of measures of oral health-related quality of life (OHRQoL) by examining the sensitivity of the oral health impact profile (OHIP)-14 to change when used to evaluate a dental care program for the elderly. *Methods:* One hundred and sixteen elderly patients attending four municipally funded dental clinics completed a copy of the OHIP-14 prior to treatment and 1 month after the completion of treatment. The post-treatment questionnaire also included a global transition judgement that assessed subjects' perceptions of change in their oral health following treatment at the clinics. Change scores were calculated by subtracting post-treatment OHIP-14 scores from pre-treatment scores. The longitudinal construct validity of these change scores were assessed by means of their association with the global transition judgements. Measures of responsiveness included effect sizes for the change scores, the minimal important difference, and Guyatt's responsiveness index. An receiver operating characteristic (ROC) curve was constructed to determine the accuracy of the change scores in predicting whether patients had improved or not as a result of the treatment. *Results:* Based on the global transition judgements, 60.2% of subjects reported improved oral health, 33.6% reported no change, and only 6.2% reported that it was a little worse. These changes are reflected in mean pre- and post-treatment OHIP-14 scores that declined from 15.8 to 11.5 ($P < 0.001$). Mean change scores showed a consistent gradient in the expected direction across categories of the global transition judgement, but differences between the groups were not significant. However, paired *t*-tests showed no significant differences in the pre- and post-treatment scores of stable subjects, but showed significant declines for subjects who reported improvement. Analysis of data from stable subjects indicated that OHIP-14 had excellent test–retest reliability with an intraclass correlation coefficient (ICC) of 0.84. Effect size based on change scores for all subjects and subgroups of subjects were small to moderate. The ROC analysis indicated that OHIP-14 change scores were not good 'diagnostic tests' of improvement. The minimal important difference for the OHIP-14 was of 5-scale points, but detecting this difference would require relatively large sample sizes. *Conclusions:* OHIP-14 appeared to be responsive to change. However, the magnitude of change that it detected in the context described here was modest, probably because it was designed primarily as a discriminative measure. The psychometric properties of the global transition judgements that often provide the 'gold standard' for responsiveness studies need to be established.

Key words: global transition judgements; oral health-related quality of life; reliability; responsiveness; validity

Dr David Locker, Faculty of Dentistry,
University of Toronto, 124, Edward Street,
Toronto, Ontario, Canada M5G 1G6
Tel: +1 416 979 4907/ext. 4490
Fax: +1 416 979 4936
e-mail: david.locker@utoronto.ca

Submitted 23 May 2002;
accepted 20 May 2003

Measures of oral health-related quality of life (OHRQoL) are beginning to be used in oral health surveys, clinical trials, and studies evaluating the outcomes of dental care programs (1, 2). They may also eventually come to play an important role in clinical practice in terms of identifying needs, selecting therapies, and monitoring patient progress (3). To date, several measures have been developed that have the potential to be used in this way (4). Although these measures are similar with respect to their conceptual basis, they differ in length, the health domains they address, and the complexity of their scoring mechanisms. In order to aid the investigator or clinician, who wishes to use a measure of OHRQoL in research or clinical practice, it is essential that the technical properties of all the measures developed to date are assessed and their performance in various contexts are described.

The first step in selecting an appropriate measure of OHRQoL is to specify measurement goals, i.e. the exact purpose in using such a measure. The goal may be descriptive, predictive, discriminative, or evaluative (5). Descriptive measures are used in population-based surveys to document the prevalence or nature of health impacts; predictive measures are used to predict a patient's health status with respect to a current or future 'gold standard' measure; discriminative measures distinguish between groups that differ in clinical condition or severity; and evaluative measures assess within-subject change occurring naturally or as a result of a clinical intervention. The second step is to identify a measure whose properties conform to the goals of the intended study. Ideally, these properties should have been verified in samples or contexts similar to those being studied (6). For example, it cannot be assumed that a measure that has proved to be reliable and valid in cross-sectional studies will necessarily be suitable for use in assessing the outcomes of clinical interventions. While cross-sectional validity and test–retest reliability are desirable properties of evaluative measures, longitudinal validity, reproducibility, and ability to detect minimally important clinical changes are their necessary properties. Guyatt et al. (7) refer to the latter property as responsiveness. To date, the responsiveness of many measures of OHRQoL has not been established. This is a significant omission, given the increasing tendency to use OHRQoL measures as outcomes in clinical trials and evaluation studies. Establishing the responsiveness of the existing OHRQoL measures would assist investigators to select the most appropriate measure, provide a basis for estimating sample sizes, and assist health professionals to interpret the meaning of changes in scores derived from the measures (8, 9). This last property is sometimes referred to as interpretability, i.e. the ability to link change scores to categories that are intuitively meaningful to clinicians (9).

There are a number of ways in which responsiveness can be assessed. One is to compare the scores on a measure prior to and following an intervention that is known to be efficacious in improving patient well-being (8). Paired *t*-tests and effect sizes can be used to determine the significance and magnitude of the change that occurs. If more than one measure is used, these tests can be used to indicate which is the most responsive.

The fact that at least two clinical trials have indicated that there is a significant and relatively substantial change in scores on the oral health impact profile (OHIP) following implant therapy (1, 2) may be taken as evidence of the responsiveness of this instrument. One of these studies used effect sizes to compare the relative responsiveness of the full 49-item version of the OHIP, the 14-item short form developed by Slade (OHIP-14; 10), and the 19-item short form developed specifically for the edentulous population (OHIP-19; 11).

A second method is to relate changes in scores over time to patients' global ratings of change in their health and well-being (12). These global transition judgements ask patients to indicate whether their health status, overall quality of life, or their component domains have improved, remained the same, or worsened over a defined period of time. Beaton et al. (6) used this approach in comparing the responsiveness of five health-status questionnaires applied to individuals with musculo-skeletal injuries. Juniper et al. (13, 14) used it to assess the responsiveness of adult and child asthma health-related quality-of-life questionnaires. As natural variations in patients' health state and variability in responses to treatment mean that some patients are likely to improve, some to remain stable, and some to deteriorate, studies using this method assess the longitudinal construct validity of change scores derived from repeat administrations of the measure (13). The global transition judgements that provide the 'gold standard' in this method may be used alone or in combination with clinical measures of change. This method has the advantage that patients who are stable can be used to assess the reproducibility of the questionnaire, while patients who change can be used to calculate minimally important differences (7, 15). The minimally important difference is defined as 'the smallest difference in score, which patients perceive as beneficial and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient's management' (16).

The main aim of this paper is to illustrate the use of this latter approach for assessing the responsiveness of measures of OHRQoL. The measure assessed was the OHIP-14 (10). Although the responsiveness of this measure has been suggested in a clinical study of implant therapy (11), we wished to use the measure to assess the outcomes of a dental care program for the low-income and institutionalized elderly. This program consisted of comprehensive care provided free of charge in a network of municipally funded clinics in the City of Toronto. The main outcome was a change in OHRQoL scores. As the clinics provide care akin to that delivered in general dental practice, the responsiveness of OHIP-14 in this context needed to be verified.

## Methods

### Subjects

Subjects were recruited when they first attended one of the four clinics that delivered the dental care program we intended to evaluate. Clients had to meet age and income requirements in order to be eligible for dental care at these clinics. All the subjects were recruited by the clinic staff, trained in the study procedures by the investigator team. The clinic staff explained the purpose of the study and the nature of the research procedures involved, after which, clients agreeing to participate were asked to sign a consent form. All survey procedures were approved by the University of Toronto's Human Subjects Certification Committee.

### Study procedures

During the first visit, the subjects were asked to complete a 22-item questionnaire designed to collect data on self-rated oral health, OHRQoL, time since last dental visit, and sociodemographic information. In order to ensure that clinic staff did not have access to the participant's responses, they were given envelopes in which to seal the completed questionnaire. One month after the completion of treatment, the clinic staff mailed a follow-up questionnaire to all those participating in the first phase. The content of this questionnaire was similar to the first, with the addition of a question to assess self-perceived change in oral health since the completion of treatment. The two questionnaires were linked by means of a unique identification code allocated to each participant. Information on each participant's dental status (dentate/edentulous) and the type of dental services received at the clinic was abstracted from the participant's dental charts.

### Measures

#### Oral health-related quality of life

Oral health-related quality of life was measured using the OHIP-14 (10). This consisted of 14 items, two from each of the seven subscales comprising the original long form of the measure. Subjects were asked, 'Over the past year, how often have you … been self-conscious … because of problems with your teeth, mouth or dentures'. Responses were scored on a simple Likert-type frequency scale with the following options and numerical codes: Always = 4, Often = 3, Sometimes = 2, Seldom = 1, and Never = 0. Scores were obtained by summing these response codes for the 14 items. Consequently, higher scores indicated worse OHRQoL. The same measure was used at follow-up, except that the OHIP items were introduced by the phrase, 'Since your last visit to the dental clinic, how often have you …'.

Subscale scores were created by summing the responses to subsets of items. Because a health domain needs to be represented by at least three or four items (17), it was not possible to use the seven domains comprising the long-form OHIP. Consequently, three domains were created that measured functional limitations/pain, psychological impacts, and social impacts. As with the original 49-item OHIP, these subscales were conceptually based rather than derived from statistical procedures such as factor analysis.

#### Global ratings of change

Participant's perceptions of change in their oral health since the completion of treatment at the clinic were assessed by a single item with a 5-point response scale ('Worsened a lot'; 'Worsened a little'; 'Stayed the same'; 'Improved a little'; 'Improved a lot'). Such transition judgements are often used as a 'gold standard' when evaluating the sensitivity to change of health-related quality of life measures (18). One advantage of these judgements appears to be that they are not affected by an individual's mood (19).

#### Other measures of oral health

At both baseline and follow-up, self-ratings of oral health were obtained from all the participants. They were asked, 'How would you describe the health of your teeth and mouth today?', and scored on a 5-point scale ranging from 'Excellent' to 'Poor'.

### Statistical analysis

#### Cross-sectional construct validity and internal consistency reliability

The cross-sectional construct validity and internal consistency reliability of the OHIP-14 when used with this elderly, low-income population was examined using pretreatment scores. The former was assessed by means of the association between pretreatment scores and participants' ratings of oral health. As the numbers in each category of the self-rating were small and the data violated the assumption of homogeneity of variances, the statistical significance of the association was determined using the Kruskal–Wallis one-way analysis of variance test. The latter was assessed using Cronbach's alpha.

*Test–retest reliability*

The pre- and post-treatment OHIP-14 scale and subscale scores of participants who reported no change in their oral health were used to assess the test–retest reliability of the questionnaire. Intraclass correlation coefficients calculated by means of a one-way random effects parallel model were calculated for this purpose.

*Longitudinal construct validity*

The longitudinal construct validity was assessed by using one-way analysis of variance to examine the association between change scores and the global transition judgements collected post-treatment. Given the method of calculating change scores, good longitudinal construct validity is indicated if those reporting deterioration have negative mean change scores, those reporting stability have change scores close to zero, and those reporting improvement have positive change scores of increasing magnitude (6).

*Responsiveness*

The responsiveness was assessed in the following ways. Change scores for the scale and subscales were calculated by subtracting post-treatment scores from pretreatment scores. Consequently, positive change scores indicate an improvement in OHRQoL, while negative scores indicate a deterioration. Effect sizes were calculated by dividing the mean of change scores by the standard deviation of the pretreatment scores. The widely used benchmarks suggested by Cohen (20) indicated the magnitude of the change observed. That is, effect sizes of 0.2 were taken to be small, 0.5 to be moderate, and 0.8 or above to be large.

Following the approach suggested by Juniper et al. (13, 14), paired *t*-tests were used to examine the significance of the within-subject change of those who changed and those who reported stability. If the measure is responsive, the former should be significant and the latter nonsignificant. These tests were performed for subjects falling into each category of the global transition judgement and for pooled data. Unpaired *t*-tests were used to compare the pooled mean change scores of those who improved with the mean change scores of those who remained stable.

The mean change scores of those reporting that they improved a little was used to determine the minimum important difference for the OHIP-14 and its three subscales (15). This value was used to calculate Guyatt's responsiveness statistic (7). Guyatt et al. (7) suggested that the most appropriate indicator of responsiveness relates the variability in test scores in stable subjects to the clinically important difference. Consequently, the index is given by the minimum important difference divided by an estimate of the within-individual variability for subjects who are stable. This denominator can be obtained from the analysis of variance table from which the ICC for stable subjects is obtained and is equal to the square root of twice the mean square error. It is identical to the standard deviation of change scores for stable subjects. The responsiveness index can be used to calculate the sample size needed for clinical trials (7).

*Change scores as 'diagnostic tests'*

Deyo & Centor (21) suggest that change scores can be considered to be 'diagnostic tests' for distinguishing between patients who improve and patients who do not. Consequently, various cut-off points can be examined with respect to their sensitivity and specificity in correctly identifying patients who improve or do not improve, using the global transition judgement as the external criterion of change. ROC curves can be constructed from these data to evaluate the diagnostic performance of the questionnaire and to allow the optimal cut-off point to be identified in terms of maximizing sensitivity and minimizing the false-positive rate. The area under the ROC curve indicates the probability of correctly identifying subjects who report improvement from randomly selected pairs of subjects who improved and who did not improve (8). Values of 0.5 and 1.00 indicate no accuracy and perfect accuracy, respectively.

*Sample-size calculations*

There are no guidelines regarding the sample size needed for a study of responsiveness (6). However, studies of responsiveness of other questionnaires have used relatively small samples, usually 40–50 subjects. Test–retest reliability assessment using the intraclass correlation coefficient, with sufficient power to detect a significant difference between the null value of 0.60 and the desired value of 0.80, requires approximately 40 subjects. Consequently, we arbitrarily set the sample size at 250 participants on the assumption that, after taking account of attrition (which we expected to be high), sufficient subjects would be available for reproducibility testing and for calculating parameters such as the minimal important difference.

## Results

### *Response*

The pretreatment questionnaire was completed by 230 subjects, of whom 128 (55.7%) also completed the post-treatment questionnaire. There were no differences between those who did and did not respond to the post-treatment questionnaire according to sociodemographic characteristics, dental status, self-rated oral health, OHIP-14 scores, time since last dental visit, pattern of dental treatment received, or the clinic at which treatment was received. A further 12 subjects were excluded from the analysis because of a high number of missing values on either the pre- or post-treatment OHIP-14.

### *Characteristics of participants*

The 116 participants included in the analysis consisted of 54 males and 62 females. Their ages ranged from 59 to 88 years, with a mean of 69.1 years. At baseline, 55.4% rated their oral health as only fair or poor and 57% reported that they had visited a private dental practitioner within the last 2 years. The proportion of participants who received each type of treatment at the clinics was as follows: examination, 99.2%; preventive, 61.1%; restorative, 50.0%; prosthodontic, 41.3%; surgical, 34.9%; other (endodontic, periodontal, …), 14.3%. Almost two-thirds (63.8%) received two or more types of dental service (preventive, restorative, prosthodontic, extractions, other) at the clinics and almost one-third (28.4%) received three or more types of treatment. For 17.2%, an examination and preventive care were the only treatments provided.

### *Cross-sectional construct validity and internal consistency reliability*

There was a significant association between pretreatment OHIP-14 scale and subscale scores and self-ratings of oral health in the expected direction ($P < 0.001$ for all analyses; Kruskal–Wallis one-way analysis of variance). Cronbach's alpha for the overall scale was 0.94, and for the three subscales, the Cronbach's alpha values were 0.81, 0.91, and 0.87, respectively.

### *Global transition judgements*

Thirty-one percent of subjects reported that their oral health was a lot more better following treatment at the clinics; 29.2% reported that it was a little better; and 33.6% reported no change. Only seven subjects (6.2%) reported that their oral health was a little worse. None of the subjects reported that their oral health was a lot worse. There was a significant association between these global transition judgements and transition categories created from the self-ratings of oral health pre- and post-treatment ($P < 0.001$). That is, the ratings of subjects who reported a worsening of their oral health declined, while the ratings of those who reported that their oral health had improved were better. The ratings of those reporting no change were stable, as indicated by an intraclass correlation coefficient (mathematically equivalent to a weighted Kappa) of 0.60.

That the majority of subjects improved is also reflected by the mean pre- and post-treatment OHIP-14 scores of 15.8 (SD = 13.7) and 11.5 (SD = 11.1; paired *t*-test, $P < 0.001$), respectively. The three subscale scores also showed a significant decline between the pre- and post-treatment administrations of the questionnaire ($P < 0.01$, in all analyses).

### *Test–retest reliability*

The 39 subjects reporting no change in their oral health status were used to assess the test–retest reliability of the OHIP-14. The intraclass correlation coefficient for the scale as a whole was 0.84 ($P < 0.0001$; 95% CI = 0.69–0.72). The ICCs for the three subscales were: functional limitation/pain = 0.82 ($P < 0.0001$; 95% CI = 0.65–0.90), psychologic impact = 0.77 ($P < 0.0001$; 95% CI = 0.56–0.89), and social impact = 0.84 ($P < 0.0001$; 95% CI = 0.68–0.91).

### *Longitudinal construct validity*

Table 1 shows mean change scores for the OHIP-14 and its three subscales for each category of the global transition judgement. All scores show a gradient from 'a little worse' (all mean scores negative) to 'a lot better' (highest positive mean change scores). However, the association was statistically significant for the social impact subscale only. When the seven subjects reporting a deterioration in their oral health were excluded from the analysis on the grounds that the small sample size means the estimate of the magnitude of change in this group is tenuous, the

Table 1. Mean OHIP-14 scale and subscale change scores by global transition judgement

| Global transition category | Number of subjects | OHIP-14 | Function/pain | Psychological impact | Social impact |
|---|---|---|---|---|---|
| A little worse | 7 | −4.00 | −2.00 | −0.43 | −1.57 |
| Same | 39 | 2.45 | 0.84 | 1.26 | 0.34 |
| A little better | 34 | 5.00 | 1.21 | 2.24 | 1.55 |
| A lot better | 35 | 7.97 | 2.14 | 2.89 | 2.97 |
| *P** | | NS | NS | NS | <0.05 |

**P*-values derived from one-way analysis of variance.

statistical significance of the associations remained the same.

## *Responsiveness*

Table 2 shows effect sizes for the OHIP-14 and its component subscales for subjects in each category of the global transition judgement. The effect size for the scale calculated using all the subjects was 0.32, and effect sizes for the subscales ranged from 0.27 to 0.34. These are estimates of the average treatment effect. Effect sizes for all those who reported that they had improved were small to moderate. The largest effect size (0.62) was found for the social impact subscale among those who improved a lot. When data for those who reported improving a little and a lot were pooled, the effect size for the OHIP-14 was 0.48 and effect sizes for the three subscales ranged from 0.41 to 0.47. Again, the small number of subjects reporting that they were a little worse means that the estimates for this group should be treated with caution.

Paired *t*-tests indicated that the difference in the pre- and post-treatment scores of those who remained stable were not significantly different. However, there was a significant difference in the pre- and post-treatment scores of those who reported improving a little ($P < 0.05$) and those who reported improving a lot ($P < 0.01$). Unpaired *t*-tests indicated that the mean change score of those who remained stable was 2.45, while the pooled mean change score for those who improved was 6.45. This difference just failed to reach statistical significance ($P = 0.09$). The small number of subjects deteriorating precluded these analyses for the group.

The minimum important difference is given by the mean change scores of those who reported improving a little. This study indicates that for the OHIP-14, with a 5-point response scale, this is equal to 5. As the standard deviation of change scores in stable subjects was 9.6, Guyatt's responsiveness statistic equals 0.54. From a table provided by Guyatt et al. (7), this means that 68 patients per group would be needed in a clinical trial to detect the minimum important difference with a one-sided test, with $\alpha = 0.05$ and $\beta = 0.20$. The number per group is relatively large as the minimum important difference is small in relation to the variability in change scores for stable subjects. Had the responsiveness statistic been 1, only 19 subjects per group would have been required.

## *Change scores as diagnostic tests of change*

When a change score of 1 or more was used to classify subjects as improved, 61.8% of those who reported improving were correctly identified. With this cut-off point, the false-positive rate was 57.8%, meaning that over half of those who reported no change or a deterioration in their oral health were incorrectly classified as having improved. Table 3 shows sensitivities and false-positive rates when a number of other cut-off points were used. When plotted for as ROC curve, the points were very close to the diagonal line. When all the data points were used to create the ROC curve, the area under the curve was 0.57, which was not significantly different from the null value of 0.5. It suggests that for any cut-off point, a subject is almost equally likely to belong to the stable or improved categories. It also indicates

Table 2. Effect sizes for the OHIP-14 and its subscales

| Global transition category | OHIP-14 | Function/pain | Psychological impact | Social impact |
|---|---|---|---|---|
| All subjects | 0.32 | 0.27 | 0.34 | 0.27 |
| A little worse | −0.29 | −0.49 | −0.08 | −0.33 |
| Same | 0.18 | 0.20 | 0.23 | 0.07 |
| A little better | 0.37 | 0.30 | 0.40 | 0.32 |
| A lot better | 0.58 | 0.52 | 0.52 | 0.62 |

Table 3. Sensitivities and 1-specificities (false positives) with different OHIP-14 change score cut-off points

| Change score cut-off point | Sensitivity (%) | 1-Specificity (false-positive rate (%)) |
|---|---|---|
| 1 | 61.8 | 57.8 |
| 2 | 55.9 | 51.1 |
| 3 | 48.5 | 48.9 |
| 4 | 44.1 | 46.7 |
| 5 | 42.6 | 40.0 |
| 6 | 40.0 | 38.2 |
| 7 | 38.2 | 33.3 |
| 8 | 35.3 | 24.4 |
| 9 | 30.9 | 24.4 |
| 10 | 22.2 | 29.4 |

Table 4. Distribution of change scores for those who improved and those who did not improve

| Change score | Not improved ($n = 46$ (%)) | Improved ($n = 69$ (%)) |
|---|---|---|
| ≤10 | 13.4 | 2.9 |
| −9 to −1 | 24.4 | 25.0 |
| 0 | 4.4 | 10.3 |
| 1–9 | 35.6 | 32.4 |
| ≥10 | 22.2 | 29.4 |

that no cut-off point is better than the other in terms of the trade-off between sensitivity and specificity. The reason for the poor performance of the change scores in predicting those who did and did not change is that in both the groups, some subjects had positive and some had negative scores (Table 4). For example, 22.2% of those who reported no change had change scores of plus 10 or more.

## Discussion

As it cannot be assumed that an OHRQoL questionnaire that has shown good discriminative properties will be suitable for use in studies to detect change (8), the responsiveness of all OHRQoL instruments developed to date should be assessed. Ideally, studies should compare the performance of two or more measures to determine which is the most appropriate, i.e. the most responsive in various research and/or treatment contexts. Such studies are relatively easy to conduct as the design is simple. It requires only that questionnaires are administered twice to a group of patients over a period of time during which some will be expected to remain stable and some to change. The second administration should ask patients for their overall perceptions of stability/change, as these judgements are used as the 'gold standard' in assessing responsiveness. The analysis of the data is also relatively straightforward because it only involves the calculation of change scores, effect sizes and minimal important differences, and the use of simple statistical tests to detect within-subject and between-group differences. Reproducibility is an essential component of responsiveness and can be assessed by statistics such as the intraclass correlation coefficient that is now a feature of most statistical packages. However, reproducibility alone does not guarantee that a measure is suitable for use in clinical trials or for evaluative studies. Theoretically at least, a measure may show good reproducibility, but be poor at detecting changes (7). Consequently, other information of the type provided here is necessary for the assessment of the usefulness of an evaluative instrument.

The analyses conducted on pretreatment data indicate that OHIP-14 had acceptable cross-sectional construct validity and internal consistency reliability when used with the elderly, low-income subjects recruited for the study. In this respect, the study confirms earlier work indicating that the measure has good discriminative properties.

However, the main aim of the analysis was to assess whether or not the OHIP-14 is a useful measure of the outcomes of what is in effect routine dental care for this elderly population. A measure is useful if it detects clinically meaningful change using feasible sample sizes (7).

The results of the study reported here are somewhat equivocal. Although mean OHIP-14 change scores showed a clear gradient in the expected direction across the categories of the global transition judgement, the differences between the change groups were not statistically significant. Moreover, although the mean change score of all those reporting improvement was more than twice that of those who reported no change, this difference was also not significant. This suggests that a larger sample size is needed in order to confirm the longitudinal construct validity of OHIP-14. In fact, in order to detect a significant difference in mean change scores between stable patients and those who improved, 96 subjects per group would have been required, which is more than twice the number included in this study. Sample-size estimates for a clinical trial to detect the minimal important difference of a 5-point change also suggest that quite large numbers of subjects would be needed in each group.

The paired *t*-tests did indicate that there was no difference in the pre- and post-treatment scores of those reporting stability, but there was a significant decline in scores for those reporting that they were a little better and a lot better. However, the effect sizes for the latter two groups were small to moderate at 0.37 and 0.58, respectively. These effect sizes are comparable to those reported by Allen & Locker (11) in a clinical trial of implant therapy. The implant group had a significant mean change score of 7.6, but this translated into an effect size of only 0.3. The 49-item version of the OHIP had an effect size of 1.0. The mean change score of the conventional denture comparison group was not significant, and represented an effect size of 0.2. The OHIP-49 did identify significant change in this group with an effect size of 0.5.

Taken together, the results of these studies do suggest that OHIP-14 is responsive to change. However, in this study, the moderate effect size for those reporting substantial improvement suggests that change may be occurring that is not being detected by the OHIP-14. The difference in the effect sizes between the long and short forms of the OHIP in the implant–conventional denture comparison also suggests that the latter is failing to detect some change that occurs. This attenuated responsiveness is likely because of the fact that short-form measures must by definition be somewhat compromised in terms of content validity. Consequently, the measure may well be modest with respect to the magnitude of change it detects, so that studies using this as an outcome will need larger sample sizes. Allen & Locker (11) and Locker & Allen (22) have demonstrated that short-form OHIPs, comprised of different subsets of items, detect more change and may be better as outcome measures in clinical trials or evaluation studies that require a shorter instrument. It should be noted that OHIP-14 showed excellent reproducibility, confirming Guyatt et al.'s (7) point that statistics such as the intraclass correlation coefficient can be misleading if used as the only indicator of responsiveness of a measure.

The ROC analysis indicated that the change scores derived from repeat administrations of the OHIP-14 were not very good at predicting which subjects reported improved oral health and which did not. The reason for this is that both groups contained individuals with both positive and negative change scores. The discrepancy between this analysis and the analysis of construct validity may be explained by the fact that, for the latter, data are analyzed at the group level. It examines whether those who improved had, on an average, higher positive change scores than those who did not improve. The ROC curve, however, is based on analysis at the individual level. That is, it examines the extent to which the change score of an individual correctly predicts his/her status in terms of improved/not improved. This suggests that while the OHIP-14 can be used in situations such as clinical trials, where the scores of groups are compared, it may not be useful in situations such as clinical practice, where individuals are the units of analysis.

That the responsiveness of OHIP-14 appears to be modest is probably because of the fact that it was developed as a discriminative measure. As we have previously argued, measures that are optimal at distinguishing between groups may not be optimal at detecting within-subject change and vice versa (22). Consequently, when developing or selecting an OHRQoL measure, an investigator must be clear with respect to measurement goals. It is essential that the content and thus the properties of a measure facilitate those specific goals.

A final point concerns the use of a global transition judgement as the 'gold standard' for evaluating change. This assumes that patients are able to judge whether or not they have changed over a period of time and also the direction and magnitude of that change (9). Consequently, most of the analyses reported here are dependent on the validity of these ratings. Some have argued that transition judgements are not valid and are more likely to be related to patients' ratings of their current state health rather than change over time (23). Accordingly, they suggest that assessments of responsiveness should not be based on retrospective methods. However, definitive conclusions about the validity of transition judgements cannot be made as no studies have investigated the psychometric properties of these indicators of change (9). Consequently, at the present time, global transition judgements represent the best option for assessing the responsiveness of health-related quality-of-life measures. The alternative – the direct estimation of treatment effects by studies involving interventions of known efficacy – requires that interventions exist that are known to lead to improvements in OHRQoL. In turn, this involves the supposition that measures of OHRQoL are available that have been shown to be responsive to clinically important change. Until the validity of global transition judgements has been demonstrated, the selection of an external indicator of change remains problematic (8). The use of additional external indicators of change such as clinician's ratings and change scores on physiologic indicators (if available) have been recommended, and would add strength to the conclusions of responsiveness studies (9).

## References

1. 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.
2. Allen PF, McMillan AS, Walshaw D. A patient-based assessment of implant stabilized and conventional complete dentures. J Prosthet Dent 2001;85:141–7.
3. Locker D. Applications of self-reported assessments of oral health outcomes. J Dent Educ 1996;60:494–500.
4. Slade G., editor. Measuring oral health and quality of life. Chapel Hill: University of North Carolina, Dental Ecology; 1997.
5. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38: 27–36.
6. Beaton D, Hogg-Johnson S, Bombardier C. Evaluating changes in health status. Reliability and responsiveness of five generic health status measures in workers with soft tissue injuries. J Clin Epidemiol 1997;50: 79–93.
7. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;2:171–8.
8. Deyo R, Patrick D. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991;12: 142S–58S.
9. Guyatt G, Osoba D, Wu A, Wyrwich K, Norman G. Methods to explain the significance of health status measures. Hamiton, Ontario: Clinical Significance Consensus Meeting Group, Unpublished paper, 2002.
10. Slade G, Slade G. Derivation and validation of a short-form oral health impact profile. Community Dent Oral Epidemiol 1997;25:284–90.
11. Allen PF, Locker D. A modified short version of the oral health impact profile for assessing health-related quality of life in edentulous patients. Int J Prosthodont 2002;15:446–50.
12. Ziebland S. Measuring changes in health status. In: Jenkinson C, editor. Measuring health and medical outcomes. London: ECL Press; 1994.
13. Juniper E, Guyatt G, Feeny D, Ferrie P, Griffith L, Townsend M. Measuring quality of life in asthma. Am Rev Respir Dis 1993;147:832–8.
14. Juniper E, Guyatt G, Feeny D, Ferrie P, Griffith L, Townsend M. Measuring quality of life in children with asthma. Qual Life Res 1996;5:35–46.
15. Juniper G, Guyatt G, Willan A, Griffith L. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994; 47:81–7.
16. Jaeschke R, Singer J, Guyatt G. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.
17. 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.
18. MacKenzie C, Charlson M, DiGioia D, Kelley K. Can the sickness impact profile measure change? An example of scale assessment. J Chronic Dis 1986;39: 429–36.
19. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. A comparison of the sensitivity to change of several health status measurements in rheumatoid arthritis. J Rheumatol 1993;20:429–36.
20. Cohen J. Statistical power analysis for the behavioural sciences, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum and Associates; 1988.
21. Deyo R, Centor R. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39: 897–906.
22. Locker D, Allen PF. Developing short form measures of oral health related quality of life. J Public Health Dent 2002;62:13–20.
23. Norman G, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869–79.

This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.