© 2007 The Authors. Journal compilation
© 2007 Blackwell Munksgaard

COMMUNITY DENTISTRY AND ORAL EPIDEMIOLOGY

Does self-weighting of items enhance the performance of an oral health-related quality of life questionnaire?

David Locker, Eva Berka, Aleksandra Jokovic and Bryan Tompson

Faculty of Dentistry, University of Toronto, Toronto, Ontario, Canada

Locker D, Berka E, Jokovic A, Tompson B. Does self-weighting of items enhance the performance of an oral health-related quality of life questionnaire? Community Dent Oral Epidemiol 2007; 35: 35–43. Journal compilation © 2007 Blackwell Munksgaard

Abstract – *Objectives:* To determine if self-weighting of the items in an oral health-related quality of life questionnaire improves its psychometric properties. *Methods:* The Surgical Orthodontic Outcome Questionnaire (SOOQ) was designed to assess the oral health-related quality of life of individuals before and after corrective surgery. Each of its 33 'items' consists of two questions: a question concerning the frequency with which a given functional or psychosocial problem had been experienced and a 'weighting' question which asked about how much the individual was bothered by that problem. The questionnaire was completed by three groups of individuals: (i) pretreatment; (ii) immediate (i.e. 2–6 months) postsurgery and (iii) postsurgery (i.e. more than 2 years after surgery). Unweighted scale scores were obtained by summing the response codes to the frequency question and weighted scores by summing the products of the frequency and bother questions. These scores were calculated for the full questionnaire and a short form consisting of 15 items. The discriminative and correlational construct validity of these scores was compared along with internal consistency reliability. The sensitivity to change and longitudinal construct validity of unweighted and weighted scores was assessed in a simulated evaluative study in which pretreatment and postsurgery subjects were paired. *Results:* For both the long and short forms of the questionnaire, unweighted and weighted scores discriminated between the groups enrolled in the study. Correlations with a general health rating were similar, as were Cronbach's alpha values and test–retest reliabilities. The simulated evaluative study suggested no differences in sensitivity to change or longitudinal construct validity. When subscale scores were examined, there was a suggestion that weighting improved their reliability. *Conclusions:* Self-weighting of items did not substantially improve the performance of the SOOQ. Domain weights should be developed and tested to determine if they have an effect on its properties.

Key words: HRQoL; OHIP; quality of life; questionnaire; self-weighting

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

Submitted 24 October 2005;
accepted 22 November 2005

The majority of health-related quality of life (HRQoL) measures consist of a sequence of items that ask about the frequency with which certain illness-related events occur. Similarly, quality of life (QoL) instruments consist of a series of items concerning different aspects of daily life and require respondents to rate their level of satisfaction with these aspects (1). Typically, questionnaire scores are obtained by summing the numerical codes attached to the categories of Likert-type response scales. This simple additive approach assumes that the events or aspects of daily life described by the items are equivalent in terms of their severity or importance to the people concerned.

This assumption of equivalence is, on the face of it, questionable. For example, the Oral Health Impact Profile (OHIP) (2) contains the following two items: 'How often in the last year have you had food trapping in your teeth or dentures?' and 'How often in the last year have you been totally unable to function…because of problems with your teeth, mouth or dentures?'. Intuitively, it seems that the latter is a more severe outcome and of more significance in terms of HRQoL than the former and that this difference should be reflected when calculating OHIP scores.

Similarly, some of the aspects of daily living described by items in QoL measures are likely to be more important to those concerned than aspects of daily living described by other items, so that a simple additive approach results in 'an inaccurate representation of quality of life' (3). Rather, item scores should be derived from some combination of satisfaction and importance ratings (1). Usually, this involves multiplying the two ratings prior to summing item scores to obtain overall or domain scores. Gill and Feinstein (4) consider that this is essential if the aim is to measure QoL rather than some other construct. As the same argument applies to the measurement of health and oral HRQoL, some instruments incorporate weights intended to reflect differences in the severity or importance of events associated with clinical conditions of various kinds (2, 5–7).

A number of different approaches to weighting have been used. Some instruments use 'set' or 'external weights' while others use self-weights applied to items or domains or both. The OHIP has set item weights developed on the basis of the views of an external panel of judges. The Thurstone method of paired comparisons was used in which each judge considered pairs of items and identified which of the pair described the more severe event. An OHIP item's score is derived by multiplying the set weight by the Likert frequency response code for that item. Scale scores are then obtained by summing the resulting values across all items.

Some have argued that this approach is inappropriate because the weights derived from the opinions of external panels are unlikely to correspond to the values, preferences or perceptions of those taking part in a study. Consequently, self-weighting approaches should be used in which each individual rates the severity or importance of each item or domain comprising the questionnaire. For example, early versions of the Dental Impact on Daily Living Scale (DIDL) used self-weightings of both items and domains (5).

The development of external weights can be a complex and time-consuming process while their use adds complexity to the calculation of scale scores. Self-weights do not require a complex development process but can increase the length and complexity of a questionnaire, or require a respondent to undertake additional procedures that may contribute to respondent fatigue and increase the probability of error. Additional computational procedures are also required that may contribute to error. Consequently, some take the view that weights should only be used if they enhance the performance of a HRQoL instrument. To date, comparisons of weighted and unweighted approaches have produced little evidence that weighting enhances the technical properties of an instrument.

For example, Allen and Locker (8) used data from a population-based study of older adults to assess the ability of the OHIP to discriminate between groups using three scoring methods; a simple count method, an additive method and a method incorporating set item weights. The analysis was undertaken for the full 49-item version and a short version consisting of 14 items. While weighted scores performed better than simple counts, they were no better than the additive scores derived from a summation of Likert response codes. A subsequent study used data from a nonrandomized trial of implant supported and conventional dentures to assess the sensitivity to change of OHIP scores obtained by the three scoring methods (9). Although the longitudinal construct validity of the scores obtained by the three methods was equivalent, effect sizes indicated that weighted scores exhibited poor sensitivity to change.

McGrath and Bedi (7) examined the contribution of item self-weights by comparing the performance of unweighted and weighted versions of the OHQoL-UK in terms of the ability of scores to discriminate between groups defined by sociodemographic and oral clinical variables. They concluded that the weighted version conferred no benefits when used in cross-sectional population studies. However, the unweighted and weighted versions were used in different studies with different samples and the scaling of the base question in the two versions differed. Furthermore, while scores from both versions showed significant associations with all independent variables, no tests were undertaken to compare the ability of the measures to discriminate between groups. Nor were test–retest reliabilities compared. This property needs to be assessed given that item self-weighting requires a respondent to answer two questions per item instead of one, and increases the possibility of random error.

Studies of item self-weights have largely been undertaken using population-based samples in which the number, frequency and severity of functional and psychosocial impacts are usually low. Whether item self-weights contribute to the performance of an instrument when used with clinical samples where the burden of oral disorders is much higher is not known (7). Moreover, the discriminative properties of measures incorporating item weights have not been fully assessed and no studies are available to determine the effect of item weights as opposed to set weights on the evaluative properties of a measure. Consequently, when developing an instrument to assess the oral HRQoL of orthodontic patients whose condition required surgical correction as part of the treatment process, items were self-weighted so that the performance of weighted and unweighted versions of the instrument could be compared. The instrument has been named the Surgical Orthodontic Outcome Questionnaire (SOOQ) (10). The aim was to develop a condition-specific instrument that provided a comprehensive assessment of the pre- and postsurgical HRQoL of patients with dentofacial conditions.

Methods

Development of the questionnaire

The content of the questionnaire was designed to reflect patients' motivations for surgical orthodontic treatment, the effects of their dentofacial condition on HRQoL and the effect of surgical orthodontic treatment on the functional and psychosocial issues encountered by these patients. Items for the questionnaire were developed based on: (i) reviews of the literature on motivations for treatment among the intended clinical group, (ii) reviews of existing oral HRQoL measures, such as the OHIP (2), the COHQoL (11) and the OQLQ (12), and (iii) expert clinical opinion. The aim in selecting/developing items was to ensure content coverage and relevance.

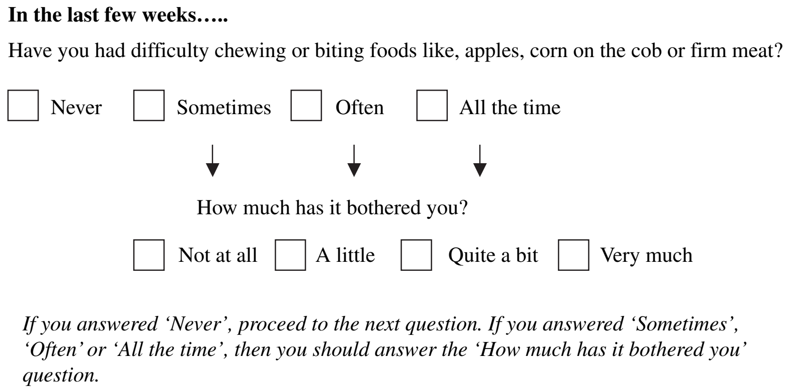
This process resulted in a questionnaire consisting of 33 items organized into five domains, namely: function 1 – issues before surgery (six items); function 2 – issues after surgery (nine items), dental aesthetics (five items), facial aesthetics (four items) and emotional and social well-being (nine items). Each item consisted of two parts: an initial question asking how frequently a given problem has been experienced and, for those reporting having experiencing the problem, a second question concerning its importance. Importance was assessed by asking participants how much the problem bothered them. The base question and the weighting question were both scored using a Likert-type response format. An example is given in Fig. 1.

The questionnaire also contained questions on self-rated general health, motivations for treatment and, for those having had surgery, satisfaction with treatment outcomes. It was formatted to facilitate self-completion according to current guidelines for self-administered instruments (13).

Evaluation of the questionnaire

Participants

Participants for this phase of the study were recruited from the Graduate Orthodontic Clinic, Faculty of Dentistry, University of Toronto and two private dental offices, one the office of a specialist in orthodontics and the other the office of a



In the last few weeks…..

Have you had difficulty chewing or biting foods like, apples, corn on the cob or firm meat?

□ Never □ Sometimes □ Often □ All the time

↓ ↓ ↓

How much has it bothered you?

□ Not at all □ A little □ Quite a bit □ Very much

If you answered 'Never', proceed to the next question. If you answered 'Sometimes', 'Often' or 'All the time', then you should answer the 'How much has it bothered you' question.

Fig. 1. Structure of the questionnaire.

specialist in oral surgery. In order to assess the discriminative properties of the questionnaire patients from three groups were recruited: (i) pretreatment; (ii) immediate (i.e. 2–6 months) postsurgery and (iii) postsurgery (i.e. more than 2 years following completion of surgery). The inclusion criteria for each group are given in Table 1. Clinic support staff identified potential participants using a checklist of inclusion criteria and questionnaires were mailed to these individuals or given to them during clinic visits. All participants were asked to sign a consent form for the study that also allowed access to charts for the abstraction of data such as initial diagnosis and type of surgery. Treatment providers were not aware which of their patients participated in the study.

All recruitment and data collection procedures and data collection instruments were approved by the Health Sciences Committee I of the Ethics Research Office, University of Toronto.

Sample size

A sample size of 30 patients per group was initially stipulated based on the work of Kiyak et al. (14). Once data collection was completed for 10 patients per group, the data were analysed and sample-size calculations undertaken. These confirmed that 30 per group was sufficient to detect significant differences in mean HRQoL scores between pretreatment and postsurgical patients. In order to assess test–retest reliability, 16 subjects in the pretreatment group completed two copies of the questionnaire. This was sufficient to detect an intraclass correlation coefficient (ICC) of 0.8 with alpha set at 0.05, beta at 0.2 and the null value at 0.4 (15). The pretreatment group was chosen for test–retest reliability assessment because patients of the immediate postsurgery group were likely to be in a state of change and those of the 2-year postsurgery group were less likely to show the variability in scores for adequate testing of reliability.

Calculation of scores

The response codes for the questions concerning frequency were; 0, Never; 1, Sometimes; 2, Often; and 3, All the time. The codes for the level of bother were: 0, Not at all; 1, A little; 2, Quite a bit; 3, Very much. Unweighted scores were obtained by summing the frequency response codes. The range of possible scores was 0–99. One set of weighted scores (weighted 1) was obtained by summing the product of the frequency and bother codes. Using this approach, those who reported that the event described by the frequency question never happened were given a score of 0. Those who reported that the event happened ‘all the time’ but were ‘not at all’ bothered by it were also given a score of 0 (3 × 0) for that question. If they were ‘very much’ bothered, their score was 9. Consequently, weighted 1 scores could range from 0 to 297. A second set of weighted scores (weighted 2) were calculated by recoding the importance question as follows, 1, Not at all; 2, A little; 3, Quite a bit; 4, Very much, and

Table 1. Inclusion criteria for the three groups

Group	Inclusion criteria
Pretreatment	Each patient had undergone a diagnostic appointment with the orthodontist and oral surgeon. He/she was aware of the nature of their condition and treatment options, including surgical intervention No treatment had been initiated
Immediate postsurgery	Each individual had undergone presurgical orthodontics and orthognathic surgery Surgery was of the following type: Le Fort I osteotomy; bilateral saggital split osteotomy, and/or genioplasty All had rigid fixation or wire fixation They were 2–6 months postsurgery All were undergoing postsurgical orthodontics
Postsurgery	Each individual had undergone presurgical orthodontics and orthognathic surgery Surgery was of the following type: Le Fort I osteotomy; bilateral saggital split osteotomy, and/or genioplasty Surgery was performed no earlier than 1996 All had rigid fixation or wire fixation All were at least 2 years postsurgery All had completed postsurgical orthodontics

summing the product of the frequency and bother codes. With this coding scheme, someone reporting that the event described by a question happened 'all the time' but were 'not at all' bothered by it would be given a score of 3 (3 × 1) for that question and if they were very much bothered, their score would be 12. Weighted 2 scores could range from 0 to 396.

Three sets of scores were also calculated for a short form of the questionnaire consisting of 15 items, three from each of the five domains. These were selected according to the item impact procedure described by Guyatt et al. (16). The ranges of short-form scores were 0–45 (unweighted), 0–135 (weighted 1) and 0–180 (weighted 2).

Assessment of technical properties of the questionnaire

Given the distribution of scores, Kruskal–Wallis tests were used to assess the significance of differences in scores between the three clinical groups in the study. Differences in mean ranks between the pretreatment and 2-year postsurgery groups were used to compare the ability of the weighted and unweighted scores to discriminate between groups (8). The correlational construct validity of the three scores was assessed by means of Spearman's rank correlations with the global rating of general health. Internal consistency reliability for both the long and short form of the questionnaire was assessed using Cronbach's alpha. Test–retest reliabilities were assessed by means of the ICC.

As this was a cross-sectional study, the evaluative properties of the three scoring methods could not be assessed directly. Consequently, an evaluative study was simulated by pairing pretreatment and postsurgery participants. Two sets of analyses were undertaken. In the first, each pretreatment subject was age-matched with a postsurgery subject. In the second, the pairs were randomly matched. Differences in pre–post treatment scores were assessed using paired *t*-tests and effect sizes calculated by dividing the difference in scores by the standard deviation of the pretreatment score. The longitudinal construct validity of change scores was assessed by Spearman's correlation coefficients with change scores derived from the pre- and post-treatment global ratings of health. In these analyses, the magnitude of statistics such as effect sizes are irrelevant; what is important is the relative size of the statistics when using unweighted and weighted scores.

Results

Participants

One hundred and eighty-three participants were identified and sent/given questionnaires. Of these, 110 completed the questionnaire. Only 95 were eligible based on the inclusion criteria; 33 in the pretreatment group, 30 in the 2–6 month postsurgery group and 32 in the ≥2 years postsurgery group. If it is assumed that all of those not completing questionnaires were eligible, this represents a response rate of 57% (95/168).

Participants ranged in age from 16 to 58 years although the majority, 85%, were under the age of 40 years. Almost two-thirds, 62.1%, were female. Two-thirds had had or were scheduled to have surgery in one jaw with one-third having or requiring surgery in both jaws. The distribution of subjects by Angles classification was: class I – 8.5%; class II – 51.0% and class III – 40.5%.

Participants' main motivations for seeking surgical orthodontic treatment were: improvement in facial aesthetics (84.4%), improvement in self-confidence (56.2%), improvement in overall oral health (53.1%), improvement in biting or chewing (50.0%), speech problems (37.5%) and temporomandibular joint problems (15.6%). Consequently, the content of the questionnaire was consistent with the main concerns of the participants with respect to their dentofacial condition.

Descriptive statistics

For each of the 'frequency' and 'bother' questions, all of the response codes were used. For example, for the question 'In the last few weeks, have you been concerned about what other people think?', the response proportions were: 'never' – 43.2%, 'sometimes' – 35.8%, 'often' – 12.6% and 'all the time' – 8.4%. Of those responding 'sometimes' or more frequently, responses to the 'bother' question were as follows: 'not at all' – 16.7%, 'a little' – 51.9%, 'quite a bit' – 20.4% and 'very much' – 11.1%. The proportions reporting that the event happened 'sometimes' or more frequently ranged from 14.7% to 81.2%. Using the weighted 1 coding scheme, mean bother ratings ranged from 1.02 to 1.94.

As expected, the ranges of mean and median values were larger for the weighted than the unweighted scores for the 33-item questionnaire (Table 2). Consequently, weighted scores identified more variability among the participants than unweighted scores. However, skewness and kurtosis

Table 2. Descriptive statistics for the long form of the questionnaire

	Unweighted	Weighted 1	Weighted 2
Range of scores	0–81	0–226	0–307
Mean (SD)	24.7 (17.1)	42.4 (46.4)	67.1 (62.8)
Median	21.0	28.0	49.0
Skewness (SD)	1.1 (0.25)	2.0 (0.25)	1.8 (0.25)
Kurtosis (SD)	1.1 (0.50)	4.2 (0.50)	3.4 (0.50)
% with minimum score	1.1	1.1	1.1
% with maximum score	0	0	0

statistics indicated that the distributions of the weighted scores were markedly non-normal. Floor effects for all scores were minimal and there were no ceiling effects. The same pattern was observed with the scores derived from the 15-item short form of the questionnaire; i.e. there was more variability in the scores but more highly skewed distributions. Ranges of scores were 0–40 (unweighted), 0–118 (weighted 1) and 0–158 (weighted 2) respectively, while skewness (SD) statistics were 0.64 (0.25); 1.72 (0.25) and 1.47 (0.24) respectively.

Cross-sectional validity and reliability

There were significant differences between the three clinical groups in the expected direction for both long-form and short-form unweighted and weighted scores (Table 3). The differences in mean ranks between the pretreatment and the 2-year postsurgery groups were the same irrespective of the form of the questionnaire and whether or not scores were unweighted or weighted. This indicates that weighting had little effect on the ability of scores to discriminate between groups. This was confirmed when effect sizes were calculated (the difference in means of the pretreatment and postsurgery groups divided by the standard deviation of the overall mean). These were; unweighted – 0.75, weighted 1 – 0.65, weighted 2 – 0.63.

The Spearman's rank correlation between unweighted scores derived from the long form of the questionnaire and the general health rating was 0.39 ($P < 0.001$). For both types of weighted scores, the correlations were 0.41 ($P < 0.001$). For the scores derived from the short form, correlations were 0.35 ($P < 0.001$), 0.38 ($P < 0.001$) and 0.38 ($P < 0.001$). Consequently, the correlational construct validity of the two forms of the questionnaires was not affected by weighting scores.

Weighting of items also had little effect on the internal consistency reliability of the questionnaires. For the long form, Cronbach's alpha values were 0.93 for unweighted scores and 0.95 for scores derived from both weighting approaches. For the short form, alpha values were 0.85 (unweighted), 0.89 (weighted 1) and 0.88 (weighted 2). Similarly, ICCs for the long form were 0.99 (unweighted) and 0.97 (weighted 1 and weighted 2). For the short form, they were 0.94, 0.94 and 0.93, respectively (for all ICCs, $P < 0.001$).

In order to further explore the effect of the number of items in a scale on the performance of unweighted and weighted scores, the cross-sectional and test–retest reliability analyses were repeated for each of the subscales comprising the SOOQ. The only difference observed was with the function 1 subscale with respect to correlational construct validity. Correlations between weighted scores and the general health rating were slightly higher than between unweighted scores and this rating (0.32 and 0.28 vs 0.20). Weighting also had some effect on the ICCs for the function 1 and function 2 subscales. For the function 1 subscale, the ICC for unweighted scores was 0.66 compared with 0.75 and 0.73 for the weighted scores. For the function 2 subscale, the values were 0.74, 0.91 and 0.87, respectively.

Tables 4 and 5 summarize the results of the two simulated evaluation studies. These show that for both the age-match and random-match analyses, the difference in pretreatment and 2-year postsurgery scores were significant. Moreover, although there is an indication that effect sizes were largest when using unweighted scores, effect sizes for weighted scores were broadly similar (Table 4). This suggests that item weights do not contribute to sensitivity to change. The correlation coefficients presented in Table 5 also indicate little effect with respect to longitudinal construct validity.

Discussion

This paper provides a comprehensive assessment of the effect of item self-weights on the performance of a measure of oral HRQoL. It confirms and extends the findings of the few studies that have addressed this issue (5, 7). Its results are also in agreement with most studies of instruments designed to measure HRQoL and QoL (1). That is, weighting of items does not appear to improve the psychometric properties of a questionnaire. This

Table 3. Discriminant validity – median scores by clinical group

	Long form			Short form		
Group	Unweighted	Weighted 1	Weighted2	Unweighted	Weighted 1	Weighted 2
Pretreatment	25	39	61	17	26	42
Immediate postsurgery	26	31	56	16	19	35
2-year post surgery	11	10	22	8	8	14
P-value[a]	<0.001	<0.01	<0.001	<0.001	<0.01	<0.001
DMR[b]	26	25	25	27	25	27

[a] *P*-values obtained from Kruskall–Wallis tests.

[b] Difference in mean ranks between pretreatment and 2-year postsurgery groups.

Table 4. Results of the simulated evaluative study: sensitivity to change as indicated by difference in pretreatment and 2-year postsurgery scores

	Age-match analysis			Random-match analysis		
	Mean difference[a]	*P*-value[b]	Effect size	Mean difference[a]	*P*-value[b]	Effect size
Long form						
Unweighted	13.1	<0.01	0.88	11.6	<0.001	0.79
Weighted 1	27.9	<0.01	0.67	23.7	<0.01	0.57
Weighted 2	40.9	<0.01	0.73	35.4	<0.01	0.63
Short form						
Unweighted	7.6	<0.001	1.05	7.2	<0.001	1.00
Weighted 1	16.9	<0.01	0.77	15.1	<0.01	0.67
Weighted 2	24.5	<0.01	0.84	22.2	<0.001	0.77

[a] Mean difference = mean of pretreatment scores − mean of 2-year postsurgery scores.

[b] *P*-values from paired *t*-tests.

Table 5. Results of the simulated evaluation study: longitudinal construct validity – rank correlations between questionnaire change scores and general health change scores

	Age-match analysis	Random-match analysis
Long form		
Unweighted	0.32	0.20
Weighted 1	0.37*	0.18
Weighted 2	0.35*	0.18
Short form		
Unweighted	0.32	0.22
Weighted 1	0.39*	0.16
Weighted 2	0.37*	0.17

*$P < 0.01$.

seems to be the case whether set weights are used or items are self-weighted. Here, the self-weighting of items did not affect the discriminative validity, correlational construct validity, or internal consistency reliability of the SOOQ. A simulated evaluative study also suggested little effect on sensitivity to change or longitudinal construct validity. While Streiner and Norman (17) have suggested that weighting may have an effect on instruments with fewer than 40 items, weighting had no effect on the performance of long (33 items) or short (15 items) versions of the SOOQ. The only noticeable effect was an increase in the range of scores and a change in their distribution to markedly non-normal. When subscale scores were examined, there was an indication that weighting improved the test–retest reliability of the two scales measuring oral functioning. However, it is probably not the case that this provides a sufficient rationale for the use of item self-weights, particularly in the light of the effect of weighting on the distribution of scores.

Given that the rationale for item weighting appears to be 'compelling' (1), the question arises as to why neither set item weights nor self-weights have any noticeable effect on the properties of a questionnaire. Allen and Locker (8) and Allen et al. (9) suggest why the set weights developed for use with the OHIP may not improve the performance of the instrument. In order to reduce the magnitude of the task for judges involved in the Thurstone paired-comparison exercise, comparisons were limited to the items in the same subscale so that the weights can be compared within but not between subscales. This resulted in a narrow range of weights; all but five of the 49 generated being below 2.0 (2). Moreover, subscale scores need to be standardized prior to calculating overall OHIP scores. This standardization contracts the range of scores resulting in a loss of sensitivity (9). This would not explain the lack of an effect in this study where weights, i.e. bother ratings, could vary between 0 and 4 (depending on the scoring approach) and where there was a considerable increase in the range of scores.

A more compelling explanation is that unweighted and weighted scores are highly correlated. Spearman's rank correlations between the three scores derived from the long version of the questionnaire ranged from 0.93 to 0.99. For the short form, they ranged from 0.94 to 0.99. Consequently, weighting had little or no effect on the ranking of individuals on the basis of their scores. This high correlation between unweighted and weighted scores has been reported by others (1).

A further consideration is that the item frequency and bother ratings were not independent. Correlations between the two ranged from 0.80 to 0.99 with the majority exceeding 0.90. This suggests that the importance of the event described by an item is a function of how often it occurs rather than a function of its intrinsic character. For example, of those who responded 'sometimes' to the item concerning food stuck in between the teeth, 8.6% reported that it bothered them 'quite a bit' or 'very much'. Of those who reported that this happened 'all the time', 82.6% reported being bothered 'quite a bit' or 'very much'. This suggests that importance ratings do not contribute much additional information over and above that derived from frequency ratings. Similar findings have been reported by Trauer and Mackinnon (1). They analysed data from a study that assessed QoL among those with mental illness. Participants were asked to rate their satisfaction with a number of aspects of daily life and also to rate the importance of those aspects. They found that overall QoL scores were 'driven' largely by the satisfaction ratings, with importance ratings contributing little additional information.

It is of course possible that the association between frequency and bother ratings is an artefact produced by the layout and design of the questionnaire. That is, responding to the frequency question may bias the response to the subsequent bother rating. Consequently, it may be worth testing alternative approaches to gathering data on the relative importance of the events described by the items in oral HRQoL questionnaires to determine if this is the case.

Nonetheless, the results of this study cast further doubt on the practice of using item self-weights when measuring oral health outcomes. In spite of the face validity of weighting items, little appears to be gained and much may be lost by increasing the length of a questionnaire in order to obtain these weights. However, given the findings of Leao and Sheiham (5) with respect to domain weights, further testing of the effects of having respondents rate the importance of domains rather than items should be explored further in both cross-sectional and longitudinal studies.

References

1. Trauer T, MacKinnon A. Why are we weighting? The role of importance ratings in quality of life measurement. Qual Life Res 2001;10:579–85.
2. Slade GD, Spencer AJ. Development and evaluation of the Oral Health Impact Profile. Community Dent Health 1994;11:3–11.
3. Ferrans CE, Powers MJ. Quality of life index: development and psychometric properties. Adv Nursing Sci 1985;8:15–24.
4. Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. JAMA 1994;272:619–26.
5. Leao A, Sheiham A. The development of a sociodental measure of dental impacts on daily living. Community Dent Health 1996;13:22–6.
6. Adulyanon S, Sheiham A. Oral impacts on daily performances. In: Slade GD, editor. Measuring oral health and quality of life. Chapel Hill, NC: University of North Carolina, Dental Ecology; 1997. p. 151–60.
7. McGrath C, Bedi R. Why are we 'weighting'. An assessment of a self-weighting approach to measuring oral health-related quality of life. Community Dent Oral Epidemiol 2004;32:19–24.
8. Allen PF, Locker D. Do item weights matter? An assessment using the oral health impact profile. Community Dent Health 1997;14:133–8.
9. Allen PF, MacMillan AS, Locker D. An assessment of the sensitivity to change of the Oral Health Impact Profile in a clinical trial. Community Dent Oral Epidemiol 2001;29:175–82.
10. Berka E. Development and initial evaluation of a new questionnaire to assess health-related quality of life before and after surgical orthodontic treatment. MSc Thesis, University of Toronto; 2004.
11. Jokovic A, Locker D, Stephens M, Kenny D, Tompson B. Validity and reliability of a measure of child oral health-related quality of life. J Dent Res 2002;81:459–63.
12. Cunningham SJ, Garratt AM, Hunt NP. Development of a condition specific quality of life measure for patients with dentofacial deformity: I Reliability of the instrument. Community Dent Oral Epidemiol 2000;28:195–201.
13. Mullin PA, Lohr KN, Bresnahan BW. Applying cognitive design principles to formatting HRQOL instruments. Qual Life Res 2000: 8:269–73.
14. Kiyak HA, McNeill RW, West RA. The emotional impact of orthognathic surgery and conventional orthodontics. Am J Orthod 1985;88:224–34.

15. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987;6:169–74.
16. Guyatt G, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. Can Med Assoc J 1986;134:889–95.
17. Streiner D, Norman G. Health measurement scales: a practical guide to their measurement and use. Oxford: Oxford University Press; 1989.

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.