CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 41

The 15-D Measure of Health Related Quality of Life: Reliability, Validity and Sensitivity of its Health State Descriptive System

Harri Sintonen
Professor,
Department of Health Policy and Management, University of Kuopio, Finalnad,
and
Visiting Fellow, National Centre for Health Program Evaluation

October, 1994
ISSN 1038-9547
ISBN 1 875677 36 4

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of General Practice and Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE receives core funding from the National Health and Medical Research Council and Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR ACKNOWLEDGMENTS

The financial support of the Academy of Finland for this study is gratefully acknowledged. My special thanks are to Markku Pekurinen for his invaluable contribution to the development of 15D in many ways. I am also indebted to Mats Brommels, Ritva Kauppinen, Timo Leino, Jouko Lönnqvist, Markku Partinen, Jukka Rautonen, Pekka Rissanen, Hannele Tikanoja and all others, who have shared either their data and/or experiences or helped otherwise in developing and testing the 15D. My special thanks are to The National Centre for Health Program Evaluation for an excellent working environment while preparing this paper and to Prosessor Jeff Richardson for valuable comments.

ABSTRACT

The 15D is a generic, 15-dimensional, standardised, self-administered measure of HRQOL, that can be used both as a profile and as a single index score measure. This paper examines the acceptability, reliability, validity and sensitivity of two versions (15D.1 and 15D.2) of its health state descriptive system as a profile measure compared with the Nottingham Health Profile (NHP), SF-20 and EuroQol by using several data sets and methods. The response and completion rates show that the acceptability is comparable to NHP, SF-20 and EuroQol. Reliability in terms of repeatability is high, even higher than for NHP. There is substantial evidence of content and construct validity (cross-sectional and longitudinal) and depression-related criterion validity. On roughly comparable dimensions the discriminatory power of 15D.1 appears to be superior to NHP, at least equivalent to SF-20, that of 15D.2 superior to EuroQol and 15D.1, and the responsiveness of 15D.1 to change seems to be similar to NHP and SF-20. The remaining 9-10 dimensions of 15D provide a large extra reserve in these respects. It is concluded that as a profile measure 15D performs equally well as NHP and SF-20, in some respects even better. The properties as a single index score measure will be explored separately.

The 15D-Measure of Health-Related Quality of Life: Reliability, Validity and Sensitivity of its Health State Descriptive System

1 Introduction

The interest in the measurement of health-related quality of life (HRQOL) has increased considerably in recent years. Disease-specific and generic (non-disease specific) measures are increasingly applied in different research contexts. Existing measures are being developed further and new ones created in a search for 'better' instruments. However, at least as far as generic measures are concerned (the class focused on here) none of the measures and approaches developed so far can claim to have established a position as **the** measure, either as a standardised system of describing health states or as a method of valuing them.

This is no wonder since the research area is very difficult. The problems start from the very lack of general agreement on the notion of HRQOL. Yet there is a broad acceptance that HRQOL is a multidimensional concept that encompasses the physical, emotional and social components associated with illness or treatment (Revicki 1989). There is also increasing recognition that HRQOL is a subjective matter and therefore how these components are affected by illness and treatment should be assessed by the individuals themselves (eg Slevin et al. 1988). How many and what dimensions should be included to represent each component, that is, how to operationalise the components as measurable dimensions, is far less agreed.

There are also other areas of disagreement. One school argues that the measurements in different dimensions should be kept separate and presented as a profile. In this way it shows where problems with HRQOL exist and possible changes in HRQOL take place. The other school (mainly economists concerned with resource allocation), although not disputing the usefulness of the profile approach, maintains that the measurements should (also) be aggregated into a single index number to obtain an overall picture of the level of HRQOL and its changes (Bullinger 1993). However, within the latter school there is further disagreement on how the measurements should be valued or weighted in the aggregation.

There are thus two major methodological problems in constructing an HRQOL measure, that produces a single index score (Culyer 1976). The first is the creation of a standardised health state descriptive system, that is, the choice of the dimensions (attributes) in which health is to be measured and the division of each dimension into discrete levels by which more or less of the attribute can be identified. The second is the valuation of the different combinations (profiles) of the levels, one from each dimension, that is, of the health states thus defined.

The purpose of this paper is to describe, how the health state descriptive system of the 15D, a 15-dimensional measure of HRQOL, was created, and to evaluate its properties as a profile measure in terms of several criteria such as reliability, validity and sensitivity. The valuation component of 15D will be described and evaluated elsewhere. The design principles of 15D are discussed in

section 2, the conceptual basis and operationalisation in section 3. The data and methods used in evaluating the properties of the 15D health state descriptive system theoretically and empirically against a set of criteria are described in section 4. The results are presented in section 5 and discussed in section 6.

2 Design Principles of 15D

The basic objective has been to develop a generic, multi-dimensional, standardised, self-administered measure of HRQOL, that could be used primarily as a single index score measure, but also as a profile measure at least in the following areas:

First, to evaluate the effectiveness and efficiency (cost-effectiveness/utility) of different health care programs and technologies within disease categories in clinical trials or in average practice and across disease categories, and to facilitate thus resource allocation decisions both at the level of clinical and health care policy. Second, in population studies to describe and quantify the HRQOL of population groups and whole populations cross-sectionally and changes in the HRQOL over time (eg to assess the need for and effect of resource reallocation between regions). Third, to assist and improve clinical practice and individual clinical decisions by pinpointing problems that need attention and to assess clinical outcomes. Finally, to describe the patient mix of various health care units (such as hospitals, health centres) and to standardise it when analysing and comparing their productivity.

In the literature several requirements have been set for a useful generic measure (Chen et al. 1975; Gilson et al. 1975; Kaplan et al. 1976; Torrance et al. 1982; Boyle & Torrance 1984; Kirsher & Guyatt 1985; Guyatt et al. 1989). These can be condensed into the following:

Feasibility and general applicability

The measure should minimise measurement burden to respondents (be brief and acceptable) and users (inexpensive to administer, easy to compute and analyse). The information needed for the measure should be possessed by the respondents without prior use of clinical, laboratory or other health services.

Reliability

The repeatability of measurements with a minimum of random error.

Validity

The degree of confidence that can be placed in the inferences drawn from the scores of a measure.

Sensitivity

The ability of the measure to distinguish between individuals and groups in different health states (discrimination) and to detect changes in individuals or groups over time (responsiveness to change in health status).

The 15D was developed to meet these criteria as far as possible. Since some of the criteria are conflicting no measure can satisfy them all completely. The purpose was to find a reasonable balance between them. The health state descriptive system of 15D is evaluated against these criteria both theoretically and empirically by using several methods and data sets.

3 Conceptual Basis and Operationalisation of 15D

Conceptually 15D is based on the results of a thorough review of how health is conceptualised in official Finnish health policy documents. Two familiar major aspects emerged: the quantity or length of life and the quality of life, ie what people's life is like regarding health and ability to perform. By subscribing to the WHO definition of health as a state of complete physical, mental and social well-being, the quality aspect was seen to be composed of three main components: biological organismic functioning (functioning of the psycho-physical system), experiential component (perceived health/illness) and social functioning in terms of ability to perform in usual roles and tasks. Two types of social functioning were of a special importance: ability to work and social participation (Sintonen 1981a).

12D

In the first 12-dimensional health state descriptive system the functioning of the psycho-physical system was operationalised in terms of 9 basic physiological functions: moving, seeing, hearing, breathing, eating, sleeping, eliminating, communicating (speaking) and mental functioning. Social functioning comprised two dimensions: ability to work and social participation. The experiential component was measured by perceived health. It was thought to capture mental health problems and felt symptoms. Each dimension was divided into 4-5 levels (Sintonen 1981a, 1981b).

15D.1

The feedback from medical profession led to a revision in 1986. Some commentators, especially psychiatrists felt that the measure was too much oriented towards physical health, neglecting, or at least dealing inadequately with the mental side. There was also some discontent with how the levels of some dimensions were worded. For these reasons, three dimensions, depression, distress and pain, were added and levels rephrased. Thus the first 15D health state descriptive system (self-administered 15D questionnaire), to be called 15D.1, was established (Sintonen & Pekurinen 1989, 1993).

Since over 20 projects have been launched where this version was/is being used. In addition, 1815 patients (aged 12-92) filled in a 15D.1 questionnaire upon arrival to primary health care centres around the country. The questionnaire included two further questions. The first was: "We have measured your health regarding the following 15 attributes (a numbered list). Considering how health in your opinion should be measured, does this list contain attributes that are not important and could therefore be omitted? Please circle the number of the attributes that could be omitted". The second was: "Considering how health in your opinion should be measured, are there any important attributes that are not in the list. If there are, could you please briefly describe what they are?" About 300 patients provided some answer. A similar survey was carried out among 1100 primary health care patients upon arrival to health centres and outpatient clinics in Helsinki.

This time about 200 patients replied to these questions.

Very few indicated attributes to be omitted. Those who did, usually listed attributes in which they themselves had no problems. They thus did not understand the question in general terms as the idea was. No clear candidate(s) for omission emerged.

Most of the respondents wanted to add something. The answers were interpreted and grouped by a public health nurse. Four major categories were found: 1) descriptions of a clinical condition or measurement (such as rheumatism, blood pressure, cholesterol level), 2) unpleasant physical symptoms (nausea, itching, dizziness, constipation etc), 3) descriptions relating to lack of vitality (lack of energy, tiredness, exhaustion, burn-out etc) and 4) mental problems (stress, nervousness, tenseness etc). There were also numerous comments relating to sexual problems.

15D.2

Based on these results, feedback on the descriptive system from many users of 15D.1, and factor analyses of empirical data in various patient groups, the descriptive system was revised in 1993. Some changes were made in the dimensions, number of levels and wording. Correlations and factor analyses consistently showed that ability to work (Q9) and social participation (Q10) were closely associated so they were merged into usual activities. A similar association was found between depression (Q13) and distress (Q14). Therefore, to make a clearer distinction between them, the contents of the latter were changed from feeling distressed and fearful to feeling anxious, stressed and nervous. Factor analyses also indicated that perceived health was a 'summary' of the other dimensions so it was replaced by a dimension of vitality (dealing with energy, tiredness and exhaustion). Also a dimension relating to sexual activity was added.

The reason for changing all 4 level scales into 5 level scales was to increase sensitivity and to make it easier for the patients to rate themselves on the scales. Level descriptions were reworded to increase clarity, completeness of content (eg bowel function was added to elimination, pain was broadened to encompass all unpleasant physical symptoms) and sensitivity, especially at the upper ('better') end of the scale. The levels of sleeping and breathing were reformulated on the advice of specialists in sleep problems and respiratory diseases, respectively. The revised questionnaire 15D.2 is in Appendix 1.

4 Evaluation of the 15D Descriptive System

Data and methods

Feasibility and general applicability is judged theoretically by considering the type of information needed to respond, how applicable the level descriptions of the dimensions are in different social and cultural settings, and measurement burden. Empirically the evaluation takes place regarding fill-up time and response and completion rates.

The reliability of the measurement scores is concerned with the degree to which they can be repeated (McDowell and Newell 1987). It is estimated by taking repeated measurements (test-retest) and determining the agreement between them. Usually this is expressed as a correlation coefficient, but Bland and Altman (1986) have shown that it is "a totally inappropiate method": it

measures the strength of a relation between two variables, not the agreement between them.

Therefore a two-step method suggested by them is used here. First, the mean and standard deviation of the differences between the test and retest scores are calculated to find out, whether the mean difference deviates significantly from zero. If it does, the data cannot be used to assess repeatability. If it does not, in the second step a repeatability coefficient is calculated, defined as the percentage of differences (cases) falling within two standard deviations from the mean difference with 95% being an acceptable standard.

The repeatability of 15D.1 and Nottingham Health Profile (NHP) (Hunt et al. 1981) was examined among patients waiting for bypass operation (n=123, aged 31-66) (Brommels 1990). After the first measurement (test), those to be operated more urgently were chosen (n=57) and the rest (n=66) remained on the waiting list. A new measurement (retest) was taken in both groups on average in three months time. Those remaining on the list were on a conservative treatment to keep their health state stable so they are suitable for a test-retest comparison. As in all analyses in this paper, the levels of each 15D dimension were scaled separately onto a 0-1 scale (1=best level, 0=dead) based on their average relative desirability elicited from a sample of Finnish population (n=243 for 15D.1 and n=213 for 15D.2) by using a continuous 0-100 ratio scale (Sintonen 1981b). The NHP scores on a 0-100 scale (0=best, 100=worst) were derived by using Finnish population weights (Koivukangas et al. 1992).

To test the reliability of responses by proxies 22 patients (average age 68, range 35-85) in the Tampere Hospice for terminal cancer treatment rated themselves on 15D.1 and their personal nurse rated the patient independently. The nurse had known the patient on average for 39 days.

Validity indicates the extent to which accurate inferences can be made based on a measure. Validation is a process of hypothesis testing, by which the degree of confidence that can be placed in the inferences to be drawn from scores on scales is determined (Streiner & Norman 1989). In this testing several aspects (types of validity) can be explored.

Content validity refers to how adequately the content of the measure reflects its aims. The question is: do all the items appear relevant to the concept being measured and are all aspects covered (McDovell & Newell 1987). A measure that includes a more representative sample of dimensions lends itself to more accurate inferences. The higher the content validity, the broader are the inferences that can be validly drawn about the person or group under a variety of conditions and circumstances (Streiner & Norman 1989).

There are no absolute standards for judging content validity. However, at least three aspects are pertinent to good content validity. Firstly, the measure should be composed of a comprehensive set of clearly defined, one-concept dimensions, each of which should make on independent and distinguishable contribution to variation in health status. Secondly, the dimensions should be relevant and socially important and capable of being affected by health care/policy (Torrance et al. 1982, Boyle and Torrance 1984). Thirdly, Kaplan et al. (1976) argue that a value component, reflecting the relative importance or goodness of the health states, is a critical element of content validity.

Content validity is evaluated here by looking, whether the process of selecting the dimensions can be expected to produce a comprehensive set of independent dimensions that adequately cover

the conceptual basis of RHQOL, and whether it allows for different views about their social relevance and importance.
McDowell and Newell (1987) suggest that factor analysis can be used to establish content validity empirically. However, we strongly agree with Kaplan et al. (1976) that factor analysis can be used (and was used here) for data reduction, that is, to find the largest number of independent (non-correlating) dimensions, not to create the construct of HRQOL. The final choice of dimensions has to be made on the basis of social relevance and importance, not on what the correlational structure of the dimensions, which is besides different in different patient groups, happens to be.

Criterion validity of a measure is examined by correlating it with another measure, ideally a 'gold standard'. If the measure correlates with the criterion measure when given at the same time, the measure shows **concurrent validity**. If the measure is able to predict a future criterion value, it shows **predictive validity** (Streiner & Norman 1989).

Since there is no gold standard, criterion validity of HRQOL measures can not be proven. However, to the extent that the Hamilton Depression Rating Scale (HDRS, Hamilton 1967) can be regarded as a gold standard for measuring depression, we can examine the depression-related concurrent validity of the mental health dimensions of 15D.1 and SF-20. The data (subsequently the depression data) are based on a six-week RCT of 209 patients in Finland. The patients enrolled were over 18 years of age and met the DMS-III-R criteria for depressive disorder with the minimum score of 16 on the HDRS. The simultaneous measurements with HDRS, 15D.1 and SF-20 took place at weeks 0, 2 and 6 (Lönnqvist et al. 1994b).

The concurrent validity was studied by correlating all simultaneous measurements of the depression and distress dimensions of 15D.1, mental health dimension of SF-20 and HDRS. A logit model was used to examine, how well the scores on all dimensions of 15D.1 and SF-20 are able to predict, whether the simultaneous HDRS score was ≤ 16 or >16.

Construct validation involves gathering external empirical evidence, convergent or discriminant, so that meaningful inferences can be made with the measure. To show convergent validity the measure should correlate highly with other variables and other measures of the same construct, to which it should correlate on theoretical grounds. Discriminant validity implies that the measure should not correlate with dissimilar, unrelated variables or measures (Streiner & Norman 1989).

To exhibit convergent evidence, extreme groups comparison with t-test is used to test the following hypotheses: 1) the elderly (65+ years) tend to have a lower mean score on each dimension than young people (17-35 years), 2) people reporting an illness or impairment tend to have a lower mean score on each dimension than people without an illness or impairment. The data consist of the combined random population samples used in eliciting valuations for 15D. The final sample sizes were 719 for 15D.1 and 1288 for 15D.2. These data sets are subsequently referred to as 15D.1 and 15D.2 valuation data, respectively.

Furthermore, t-tests are used to test the following hypotheses: 1) at baseline the depression patients show significantly lower scores than the general adult population at least in depression, distress, perceived health, sleeping, mental function, working and social participation, 2) at baseline the patients waiting for a bypass operation report significantly lower scores than the adult population at least in mobility, breathing, pain, sleeping, working, social participation, depression and distress, and 3) bypass patients report significantly lower scores than depression patients in

mobility, breathing, and pain, whereas the opposite applies to depression, distress, sleeping and mental function. These hypotheses are based on well-known attributes of these patient groups.

Multitrait-multimethod matrix (Campbell & Fiske 1959) is used to look at convergent and discriminant validity simultaneously. It is expected that the Pearson correlations of 15D with comparable 15D and NHP, SF-20 and EuroQol dimensions are higher (convergent validity) than with non-comparable ones (discriminant validity). When comparing 15D.1 and NHP we use the bypass follow-up data (all measurements). The comparison of 15D.1 and SF-20 is based on the depression data (all measurements), and that of 15D.2 and EuroQol on a random population sample of 500 (aged 17-91, response rate 72%, subsequently called the 500 data).

Sensitivity of a measure entails two aspects. First, the ability to distinguish between individuals and groups in different health states cross-sectionally (discriminatory power) and second, to detect changes in individuals or groups over time (responsiveness to change in health status) (Kirshner & Guyatt 1985).

Patrick and Erickson (1993) mention three criteria for evaluating discriminatory power. First, the ability to detect health problems, especially in a relatively healthy population. Second, the ability to detect improving health among quite healthy people, and to avoid the 'ceiling' effect of having no better health state to go to. Third, the ability to detect worsening health among people who are already quite ill, that is, to avoid the 'floor' effect of having no worse health state to go to.

Sensitivity is evaluated theoretically by considering factors that should contribute to sensitivity. The discriminatory power is examined empirically by looking at the percentages of respondents in various patient and population groups that score the 'ceiling' for different dimensions and the measure as a whole. The corresponding percentages at the 'floor' indicate the range of health states used. The skewness coefficient reflects the skewness of frequency distributions of scores and thus sensitivity over the range. A coefficient value of 0 indicates a normal distribution. The more negative or positive the coefficient is, the more skewed the distribution is to the left or right, respectively.

Responsiveness to change was explored by comparing the percentages at the 'ceiling' and 'floor' and skewness coefficients in two patient groups at baseline and after treatment. In addition two measures of responsiveness were calculated for different dimensions: the effect size and standardised mean response (SMR). Effect size is defined as (change in the mean score from baseline to follow-up)/(standard deviation at baseline) (Kazis et al. 1989). Cohen (1977) regards an effect size of .20-.49 as small, .50-.79 as moderate and >.80 as large. The SMR is (mean response)/(standard deviation of responses), which equals the paired t-statistic without sample size factor (Liang et al. 1990).

Sensitivity of 15D.1 vs. NHP was analysed by using the bypass follow-up data. 'Ceiling' and 'floor' effects and skewness were defined for the whole sample at baseline and at one year follow-up, effect size and SMR for those operated. The comparison of 15D.1 and SF-20 is based on the depression data (baseline vs. 6 weeks). Discriminatory power of 15D.2 versus EuroQol in a relatively healthy population was evaluated by using the 500 data. A corresponding comparison between 15D.1 and 15D.2 was carried out by using the 15D.1 and 15D.2 valuation data, respectively. The samples were made comparable and compatible with the age and gender structure of the whole adult population (Statistics Finland 1993) by appropriate weighting.

5 Results

Feasibility and general applicability

Theoretical considerations

Answering the questionnaire does not require any special information - the information needed is possessed by the respondents without prior use of clinical, laboratory or other health services. The level descriptions have been phrased to minimise the effect of people's varying external social and environmental circumstances and socially differentiated sex roles on the rating. Therefore the measure is well suited for populations in different cultural and social settings.

The questionnaire is brief and designed to be self-administered. It should not impose any major burden upon respondents. If wanted or required, there should not be any major problems with interviewer-administration. If the person is for physical or mental reasons unable to reply in either way (can not read or see, confused etc), the questionnaire can be filled up for him/her by someone who knows the person well.

Empirical evidence: response and completion rates

It usually takes 5-10 minutes to fill in the 15D questionnaire. An extensive experience shows that 15D is well received and accepted. This is also reflected in response and completion rates. In postal patient or population surveys in Finland the range of response rates has been 65-80% depending on whether reminders have been mailed and what else the questionnaire includes (Pekurinen et al. 1991; Apajasalo et al. 1994; Rissanen, unpublished data).

The response rate is also very good in long-term follow-ups. For example, breast cancer patients on two drugs have been followed-up in an RCT until disease progression. They have filled in a 15D.1 questionnaire at two months intervals. During the first six years of follow-up 13.9% of measurements were missing (study in progress).

In the one-year follow-up of bypass patients, the completion rate by dimensions was 99-100% for 15D.1 and 93-97% for NHP among those, who did not drop out. In the depression data, the corresponding rates were 96-97% for 15D.1 and 97-98% for SF-20. In a population sample of 2000, the rates for 15D.1 were 97-99% (Rissanen, unpublished data). In the 500 data, the completion rates were 96-99% for 15D.2 (except sexual activity 90%) and 97-98% for EuroQol. Regarding response and completion rates, the acceptability is thus comparable to NHP, SF-20 and EuroQol. The main reasons for non-response or non-completion are probably not measurement burden or unacceptability of the questions. The lower completion rate for the dimension of sexual activity may indicate that this dimension is slightly less acceptable than the others.

However, a small number of missing responses on any dimension is not a problem, since they can be predicted with a great accuracy. For example, in the 500 data, the level rated by the respondent on any dimension was correctly predicted on average in 80% of cases (range 67-98%) with a

regression model with the level rated as the dependent variable and the levels rated on the other dimensions, age and gender as explanatory variables. It appeared that a more parsimonious model would do almost equally well so a missing response for a few dimensions can be predicted quite accurately.

Reliability

Among the patients waiting for a bypass operation the mean difference between the test and retest scores at three months varied from -0.05 to 0.03 on the 15D.1 dimensions and from -3.2 to 6.5 on the NHP dimensions. None of the differences were significantly different from zero (not even at a 10% level). By dimensions, the percentage of cases lying within two standard deviations from the mean difference was 92-100% for 15D.1 and 89-95 for NHP. Thus the repeatibility of 15D compares equally, even favourably with that of the NHP, being quite high for both measures in this data set.

The agreement between the responses of cancer patients and their personal nurses varied depending on the dimension. For the dimension "working", the agreement was 100%. On the other hand, the nurse ratings were significantly ($p < 0.05$) better for seeing, eating, elimination and mental functioning and worse for depression and distress. For many dimensions the mean differences showed great variation (sd > average distance between two levels). Partly these results reflect the small sample size. Overall the agreement was not very good, tending to be lowest, as one could expect, for the subjective dimensions. This is in line with earlier findings (eg Epstein et al. 1989, Slevin et al. 1988) emphasising the fact that HRQOL is a subjective matter and should be assessed subjectively by the individuals themselves if possible. Perhaps a close family member might be able to give more reliable responses.

Content validity

Theoretical considerations

The 15D covers the physical, psychological and social aspects of health as defined by the WHO and widely regarded as a conceptual basis for RHQOL measures. When operationalising them into 15 measurable dimensions and their levels (15D.2), health politicians, physicians, patients, other researchers and empirical research data have had a fair say. The dimensions of 15D.2 are also in concordance with those suggested by Fallowfield (1990) based on a broad analysis of literary, philosophical and scientific sources. In most cultures unimpaired basic physiological functions and other dimensions of 15D are characteristic and relevant for an individual to be regarded as healthy. Moreover, the dimensions are such that they can be affected at least to some extent by health care/policy. Kaplan et al. (1976) argue that not accounting for symptoms is a major sacrifice of content validity - in 15D.2 they are accounted for. Moreover, 15D includes a value component (to be discussed elsewhere), reflecting the relative importance or goodness of the health states as experienced by the general public.

Empirical evidence

In extensive patient surveys, no clear dimensions to be omitted emerged. The most frequently

suggested missing attributes were added. The high completion rates indicate that the dimensions appear relevant so the respondents do not object to or omit them. The 15D.2 includes the same dimensions as the other well-known measures of a similar type do - and much more (see, eg Lovatt 1992; Rosser 1993) being thus more comprehensive in sampling items for the construct of HRQOL. In the light of theoretical and empirical evidence and relative to many other measures the content validity of 15D is assuring.

Criterion validity

The correlations of the depression and distress scores of 15D.1 with the HDRS score were -.62 and -.59, respectively. When the scores were summed, the correlation was -.64. The correlation between the mental health score of SF-20 and the HDRS score was -.73. In 77% of measurements, the scores on all 15D.1 dimensions were able to predict correctly, whether the simultaneous HDRS score was ≤16 or >16. For SF-20 this was 81%. These figures provide substantial evidence for the depression-related criterion validity of 15D.1 also longitudinally.

Construct validity

The information of three multitrait-multimethod matrices has been condensed into Table 1. For example, in the column entitled NHP the correlations of 15D dimensions with comparable NHP dimensions are in bold print, and the range of correlations with non-comparable ones, either on 15D or NHP, are in parentheses. Table 1 shows that the correlations of 15D with comparable NHP, SF-20 and EuroQol dimensions are consistently higher than the correlations with non-comparable scales measuring dissimilar attributes. This is a pattern that scales with convergent and discriminant validity are expected to exhibit. In general, the correlations of 15D dimensions with comparable SF-20 and EuroQol dimensions are higher than with comparable NHP dimensions. Correlations of a similar magnitude and pattern between comparable 15D.1 and NHP dimensions were observed by Rissanen et al. (1994) among hip and knee replacement patients. These findings provide solid convergent and discriminant evidence for the construct validity of 15D.

Table 1
Multitrait-multimethod matrix of correlations of 15D dimensions (.1=15D.1 and .2=15D.2) with comparable NHP, SF-20 and EuroQol dimensions (the range of correlations with the non-comparable ones in parentheses in absolute values)

15D dimensions	15D.1 vs NHP	15D.1 vs SF-20	15D.2 vs EUROQOL
Mobility.1 [Mobility.2]	PM/N: **-.49** (.22-.43)	PF/S: **.53** (.14-.41)	[MO/E: **-.77**] (.17-71)
Sleeping.1	S/N: **-.68** (.32-.44)		
Pain.1 [Discomfort/Symptoms.2]	P/N: **-.51** (.30-.43)	P/S: **.73** (.15-.43)	[PD/E: **-.65**] (.31-.46)
Depression.1 [Depression.2]	EM/N: **-.60** (.29-.44ϖ)	MH/S: **.77** (.15-.66ϖϖ)	[MD/E: **-.67**] (.13-.36ϖϖϖ)
Distress.1 [Distress.2]	EM/N: **-.53** (.24-.40ϖ)	MH/S: **.70** (.16-.58ϖϖ)	[MD/E: **-.58**] (.13-.43ϖϖϖ)
Working.1		RF/S: **.69** (.31-.68)	
Social participation.1		SF/S: **.74** (.25-.68)	
Perceived health.1		HP/S: **.67** (.35-.66)	
[Usual activities.2]			[UA/E: **.76**] (.24-.65)

Notes

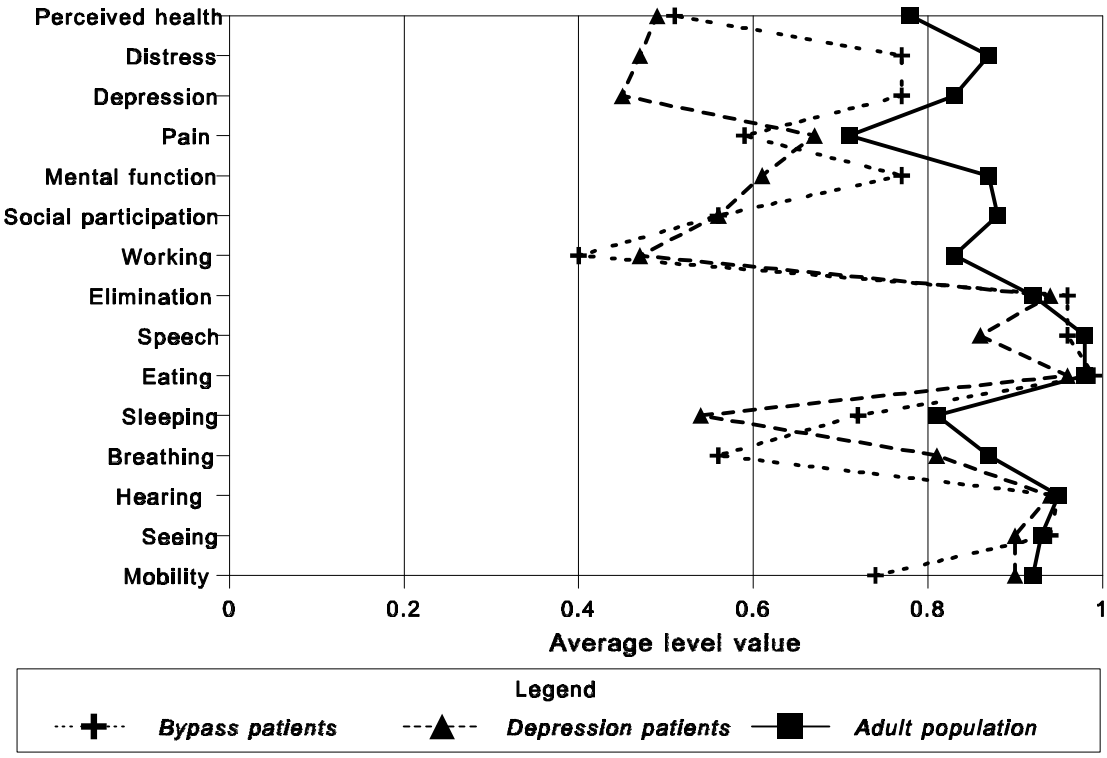
ϖ	Correlation between depression.1 and distress.1 = .70.
ϖϖ	Correlation between depression.1 and distress.1 = .76.
ϖϖϖ	Correlation between depression.2 and distress.2 = .63.
NHP	PM/N = physical mobility, S/N = sleep, P/N = pain, EM/N = emotional reactions.
SF-20	PF/S = physical functioning, RF/S = role functioning, MH/S = mental health, HP/S = health perceptions, P/S=pain, SF/S=social functioning.
EUROQOL	MO/E=mobility, UA/E=usual activities, PD/E=pain or discomfort, MD/E=mood.

The extreme group comparisons showed that apart from speech (communicating), depression and distress, the elderly (65+ years) had a lower ($p<.01$) mean score on each dimension of 15D.1 than young people (17-35 years). On 15D.2 the elderly (65+ years) had a lower ($p<.001$) mean score on each dimension except depression ($p=.11$). People reporting an illness or impairment had a lower ($p<.001$) mean score on each dimension of 15D.1 except speech ($p=.013$) and depression ($p=.008$) than people without an illness or impairment. On 15D.2 the difference was significant ($p<.001$) on each dimension. Thus the results to a great extent support our hypotheses and provide thus substantial convergent evidence of construct validity.

Still further evidence can be derived from the fact that all the expected differences between the general population and depression and bypass patients as well as between these patient groups

were confirmed. This allows distinctive profiles be created for different patient groups compared with the population. As an example, the profiles for these groups are depicted in Figure 1.

Figure 1
The profiles of adult population, depression and bypass patients on 15D.1



Sensitivity

Theoretical considerations

Patrick and Erickson (1993) suggest that to increase the discriminatory power and responsiveness to change (for better) among quite healthy and to avoid the 'ceiling' effect the measure should include dimensions like emotional well-being, positive affect, vitality and health perceptions. The 15D.1 includes emotional well-being (depression, distress) and perceived health, and the 15D.2 depression, distress, vitality and symptoms/discomfort. These dimensions should help to discriminate between, and measure change in relatively well populations. The lower levels of the 15D dimensions like mobility, breathing, eating, elimination and mental function should guarantee the ability to detect worsening health among people who are already quite ill and thus to avoid the 'floor' effect even in a very frail elderly population.

Theoretically, 15D.1 defines 10 billion and 15D.2 over 30.5 billion mutually exclusive health states (plus unconscious and dead). There is thus a great potential for discriminatory power and responsiveness to change (an enormous number of states to be in and to go to). Although most of them may never occur in practice, still regarding the number of health states defined, both versions are more sensitive than most comparable instruments producing a single index score. For example Rosser/Kind defines 28 states (Rosser & Kind 1978), EuroQol 243 states (Sintonen

1993), GWB 1548 states (Kaplan & Anderson 1988) and McMaster 960 states (Torrance et al. 1982).

Empirical evidence: discriminatory power by dimensions

Table 2 shows the discriminatory power of 15D.1 vs. NHP and SF-20, and of 15D.2 vs. EuroQol by comparable dimensions. Apart from pain and possibly mobility, 15D.1 and NHP show similar percentages of patients at the 'ceiling' and 'floor' both at baseline and end. The pain dimension of NHP does not seem to detect mild pain. The differences in percentages for mobility may reflect different contents of the dimensions: physical mobility/N reflects also self-care and body movements, not just ability to walk as mobility.1. The frequency distributions of the NHP responses were much more skewed than those of the 15D.1 indicating thus less sensitivity over the range. Moreover, at the one-year follow-up, the skewness of the NHP responses had increased considerably. These findings obviously reflect a lack of discriminatory power for the NHP.

Table 2

The discriminatory power in terms of `ceiling' and `floor' effects and skewness coefficient, and responsiveness to change in terms of effect size and standardised mean response of comparable 15D (.1=15D.1 and .2=15D.2), NHP (/N), SF-20 (/S) and EuroQol (/E) dimensions

Data and dimensions	At ceiling %		At floor %		Skewness		Effect size	Standard mean response
	Baseline	End	Baseline	End	Baseline	End		
Bypass								
Mobility 1	25.2	59.2	0.8	0.0	0.2	1.0	1.1	1.1
Physical mobility/N	20.0	44.8	7.0		0.7	1.7	0.8	0.9
Sleeping 1	22.0	28.0	2.5	0.0	0.5	0.5	0.5	0.5
Sleep/N	21.0	40.7	4.2	2.7	0.7	1.4	0.5	0.6
Pain 1	2.2	20.0	1.6		0.7	0.2	0.8	0.7
Pain/N	22.2	47.7	2.6	0.0	1.0	1.4	0.7	0.8
Depression 1	22.1	41.1	1.6	0.0	1.0	0.6	0.4	0.4
Distress 1	24.4	44.6	2.5	0.0	0.4	0.8	0.4	0.4
Emotional reactions/N	25.4	37.0	1.8	2.8	1.1	1.6	0.4	0.4
Depression								
Mobility 1	74.0	80.6	1.0		1.0	1.8	0.1	0.1
Physical functioning/S	25.6	44.0	4.2	1.8	0.5	1.2	0.4	0.5
Working 1	8.7	24.5	16.0	7.5	0.2	0.2	0.5	0.5
Role functioning/S	15.8	41.2	66.5	46.1	1.2	0.1	0.5	0.5
Depression 1	1.0	10.0	11.7	2.5	0.5	0.5	1.0	0.8
Mental health/S	0.5				0.6	0.2	1.4	1.1
Perceived health 1	2.0	11.2	4.2	0.6	0.2	0.2	0.7	0.7
Health perceptions/S	2.0	0.6			1.1	0.8	1.0	0.7
Pain 1	21.8	21.0	2.4	0.6	0.4	0.4	0.2	0.2
Pain/S	14.4	28.1			0.4	0.2	0.5	0.5
Social participation 1	21.8	21.2	2.4	2.8	0.1	0.2	0.2	0.2
Social functioning/S	12.6	25.6			0.7	0.2	0.7	0.6
Population								
Mobility 2	82.2		2.2					
Mobility/E	78.8	0.2	1.5					
Usual activities 2	62.4	1.0	2.1					
Usual activities/E	70.1	1.7	1.8					
Depression 2	61.1	1.2	1.0					
Distress 2	62.6	0.8	1.2					
Mood/E	85.1	1.1	2.4					
Discomfort/symptoms 2	40.4	0.4	0.0					
Pain or discomfort/E	58.2	1.8	0.6					

The differences in percentages at the 'ceiling' and 'floor' are a little more marked between 15D.1 and SF-20 (Table 2). The differences in percentages may reflect more differences in the contents of the dimensions than in discriminatory power. The difference in contents is most marked for mobility: mobility.1 measures ability to walk, not also body movements and self-care as physical functioning/S, which are at least partly picked up by some of the remaining 15D dimensions. Role functioning/S obviously suffers from a lack of sensitivity. Apart from physical functioning/S and pain/S the distributions of the SF-20 responses at baseline were more skewed than those of the 15D.1. At the end of the follow-up, the skewness was comparable to 15D.1. Apart from mobility, the much higher percentages at the 'ceiling' for EuroQol suggest that it has less discriminatory power than comparable 15D.2 dimensions.

Table 3 shows the discriminatory power of 15D.1 vs. 15D.2. The percentages at the floor are quite similar, but those at the ceiling are usually clearly lower for 15D.2. With a few exceptions, also the frequency distributions of 15D.2 responses are less skewed. This suggests that we have succeeded with the 15D.2 in the aim of increasing the sensitivity of some scales, especially at their upper ('better') end.

It thus appears that on roughly comparable dimensions, 15D.1 is superior to NHP and about equivalent to SF-20 in discriminatory power. The 15D.2 shows much greater discrimination than EuroQol. Moreover, it must be borne in mind that the remaining 9-10 dimensions of 15D provide an extra sensitivity reserve both at the ceiling and floor (see the discriminatory power of the measure as a whole below). The 15D.2 seems to possess greater discrimination than 15D.1.

Discriminatory power of the measure as a whole

Among the patients waiting for bypass operation, NHP located 6.5% of the patients at the 'ceiling', that is 'healthy', 15D none ('healthy'= at the best level of each dimension). None was at the 'floor' on either measure. At one-year follow-up (including those operated and those still on the list), the corresponding proportions were 13.9% and 7.6%. Thus the discriminatory power of 15D.1 seems better than that of the NHP.

Among the depression patients, neither 15D nor SF-20 detected a single 'healthy' or 'floor' patient at baseline or at two weeks of treatment. After the trial (week 6) 15D detected one 'healthy' patient (1/159), SF-20 none. The discriminatory power of these measures is similar in this patient group.

Table 3
The discriminatory power of comparable 15D.1 and 15D.2 dimensions
in two representative population samples

Dimensions	At ceiling %		At floor %		Skewness	
	15D.1	15D.2	15D.1	15D.2	15D.1	15D.2
Mobility	85.5	80.7	0.7	0.2	-2.9	-2.3
Vision	88.4	78.5	0.3	0.3	-3.0	-2.7
Hearing	91.5	82.6	0.2	0.1	-3.5	-2.6
Breathing	74.6	69.6	1.4	0.7	-1.8	-1.5
Sleeping	47.9	48.1	0.5	0.3	-0.6	-1.1
Eating	96.9	96.7	0.2	0.1	-6.6	-6.7
Speech	95.9	89.4	0.2	0.1	-5.3	-3.4
Elimination	84.4	73.3	0.4	0.2	-2.4	-1.7
Mental function	72.5	73.2	0.5	0.2	-1.4	-1.5
Depression	38.3	53.6	1.0	0.6	-1.0	-1.4
Distress	59.4	56.2	1.0	0.5	-1.3	-1.0
Pain.1/Discomfort and symptoms.2	27.0	40.2	2.1	0.4	-0.6	-0.9
Working.1/Usual activities.2	68.0	71.2	2.2	0.9	-1.9	-1.9
Social participation.1/Usual activities.2	81.6	71.2	1.6	0.9	-2.5	-1.9
Vitality		46.4		0.6		-1.3
Sexual activity		73.9		2.4		-1.9

In a population sample of 500, EuroQol classified 51.6% as 'healthy', 15D.2 only 20.7% (none were at the 'floor' on either method). Thus the discriminatory power of 15D.2 is much better than that of EuroQol in the general public.

Responsiveness to change

Table 2 shows that in the light of the effect size and SMR, 15D.1 is roughly comparable to NHP and SF-20 in the responsiveness to change.

6 Discussion

The goal has been to develop a generic, multi-dimensional, standardised, self-administered measure of HRQOL, that would meet the criteria of acceptability and general applicability, reliability, validity and sensitivity, and could be used primarily as a single index score measure, but also as a profile measure for several purposes. The development work has resulted in two 15-dimensional versions 15D.1 and 15D.2. This paper has focussed on their properties as a profile measure compared with the Nottingham Health Profile and SF-20, both explicitly designed as profile measures, and with EuroQol, which is primarily a single index score measure.

The empirical evidence suggests that both versions are very well received and accepted. Both versions take 5-10 minutes to complete. In postal patient or population surveys the response rates have varied from 65 to 80%. However, besides the 15D questionnaire, these studies have included a lot of other questions as well or have not mailed reminders. With a 15D questionnaire alone and two reminders a response rate of well over 80% can be expected at least in Finland. Unfortunately, such a survey has not yet been carried out. Also the response rate in long-term follow-up has proved to be high. Of course, in other countries, the response rates may be different.

The completion rates by dimensions have been very high for both versions. For 15D.1 the rates have varied from 96 to 100%, being equivalent to SF-20 and slightly better than for NHP. A similar range was found for 15D.2 with a slight exception, as one might expect, of sexual activity. Yet the inclusion of that dimension is both theoretically and empirically well grounded. Moreover, missing values on any dimension can be accurately predicted with the scores on other dimensions by using a regression model.

The test-retest repeatability of 15D.1 was quite high, even higher than that of NHP. This result was obtained among patients waiting for a bypass operation with the interval between the measurements being about three months. The patients were on a conservative treatment to keep their status stable. Yet three months is a so long period that changes in the patients' health status may have occurred on some dimensions thus possibly detracting from reliability. With a customary 1-2 week interval, the repeatability might have been even better. Unfortunately, empirical evidence on the repeatability of 15D.2 is not yet available, but producing it soon is solidly on agenda. Meanwhile, given the process by which 15D.2 was developed from 15D.1 there is no reason to expect that the reliability of 15D.2 would be inferior to that of 15D.1, rather on the contrary.

The results also provide substantial evidence of several types of validity. As to content validity, 15D is more comprehensive in sampling dimensions for the construct of HRQOL than most other well-known measures of a similar type. Some of the dimensions are such that at least serious problems on them are rare (eg eating, seeing). This means that they correlate poorly with other dimensions and contribute little to total variation and would therefore be omitted, if the choice of dimensions is based purely, eg on factor analysis. Therefore someone might question their inclusion. Yet these dimensions may be socially highly relevant and important and when problems on them occur they may have a profound effect on HRQOL. In this way they provide extra sensitivity for the measure. Operationalising the core of the measure in terms of basic physiological functions seems to add to the content and face validity in a clinical sense for physicians, since signs and symptoms in these functions are routinely examined by them.

The multitrait-multimethod matrix provided clear convergent and discriminant evidence of

construct validity. The convergent validity correlations of 15D were higher with comparable SF-20 dimensions than NHP dimensions suggesting that 15D may be more closely related to the former. In general, these validity coefficients were quite high being between .49-and .77. In their review of health measures McDowell and Newell (1987) found that the validity coefficients fell typically between .20 and .60. The extreme group comparisons and the ability of 15D.1 to discriminate in a predictable way between patient groups and patient groups from general population provided further evidence of cross-sectional clinical construct validity. The latter feature allows distinctive profiles be created for different patient groups compared with the general population. Such profiles are available for several patient groups as well as standards for various population groups for both versions of 15D.

Theoretical and empirical evidence also suggest that 15D has a good discriminatory power and responsiveness to change. In discriminatory power on roughly comparable dimensions 15D.1 appears to be superior to NHP and at least equivalent to SF-20, and 15D.2 superior to EuroQol and 15D.1. In fairness to EuroQol, it was never designed to be a stand-alone measure, let alone profile measure, but a simple linkage tool between more comprehensive measures (EuroQol Group 1990). In that capacity it can be quite useful, since it takes less than 1 minute to complete and is far less demanding than any other measures considered here. On comparable dimensions the responsiveness of 15D.1 to change seems similar to NHP and SF-20, providing thus also evidence for longitudinal construct validity. In addition it has to be borne in mind that in comparison to these three measures, the remaining 9-10 dimensions of 15D add the comprehensiveness of content and provide a huge extra reserve of discriminatory power and responsiveness to change.

It has been emphasised that reliability and discriminatory power may be conflicting features (Streiner & Norman 1989) as may be discriminatory power and responsiveness to change (Kirshner & Guyatt 1985). It has been particularly encouraging to learn that the 15D seems to meet these properties to a high degree simultaneously. These features are necessary in most areas, where 15D is thought to be used, especially in health program and technology evaluations.

Even if 15D is thought to be primarily a single index score measure (its properties in this capacity will be explored elsewhere), 15D.1 seems to perform as a profile measure equally well as purpose-built profile measures NHP and SF-20, in some respects even better. For obvious reasons there are more empirical results on the older 15D.1 version than 15D.2. However, bearing in mind the multi-stage process and the experiences of numerous users over several years that were used when developing 15D.2 from 15D.1 there are all reasons to assume that 15D.2 is even better in reliability and validity; in sensitivity it has already been shown to be better. Data are being collected to substantiate this assumption. Anyway the author will use the new version (15D.2) in future studies and advises the other to do the same.

15D has already proven to be a useful tool in measuring effectiveness or utility of medical interventions (eg Rissanen et al. 1994; Sintonen et al. 1994), in assisting and improving clinical practice and individual clinical decisions (Markku Partinen, personal communication) and population surveys. It is hoped that this paper will encourage its wider use in various areas.[1]

[1] When writing this, the 15D.2 questionnaire is available in English, Finnish and Norwegian. Translations into German, Japanese and Swedish are in progress.

REFERENCES

Apajasalo M, Sintonen H, Siimes MA, Hovi L, Holmberg C, Boyd H, Mäkelä A, Rautonen J. (1994) Health-related quality of life of adults surviving malignancies in childhood. (under editorial consideration).

Bland JM, Altman DG. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, Feb. 8, 307-310.

Boyle MH, Torrance GW. (1984) Developing multiattribute health indexes. Medical Care 22, 1045-1057.

Brommels M. (1990) Assessing coronary artery bypass surgery: survival, physical ability and quality of life. In ISTACH. Abstracts of the Sixth Annual Meeting, Houston, Texas, USA, (May 20-23), 8.

Bullinger M. (1993) Indices versus profiles - advantages and disadvantages. In Walker SR, Rosser RM. (Eds.) Quality of life assessment. Key issues in the 1990s. Kluwer, Dordrecht, 209-220.

Campbell DT, Fiske DW. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bulletin 56, 81-105.

Chen MM, Bush JW, Patrick DL. (1975) Social indicators for health planning and policy analysis. Policy Sciences 6, 71-89.

Cohen J (1977) Statistical power analysis for the behavioral sciences. Academic Press, New York.

Culyer AJ. (1976) Need and the National Health Service. Economics and social choice. Martin Robertson, London.

Epstein AM, Hall JA, Tognetti J, Son LH, Conant L. (1989) Using proxies to evaluate quality of life. Med Care 27 (Suppl), S91-98.

The EuroQol Group (1990) EuroQol: a new facility for the measurement of health-related quality of life. Health Policy 16, 199-208.

Fallowfield L. (1990) The quality of life. The missing measurement in health care. Souvenir Press, London.

Gilson BS, Gilson JS et al. (1975) The sickness impact profile. Development of an outcome measure of health care. Amer. J. of Public Health 65, 1304-1310.

Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A (1989) Responsiveness and validity in health status measurement: a clarification. Journal of Clinical Epidemiology 42, 403-408.

Hamilton M. (1967) Development of a rating scale for primary depressive illness. British Journal of Social and Clinical Psychology 6, 278-296.

Hunt SM, McKenna SP, McEven J, Williams J, Papp E. (1981) The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med 15A, 221-229.

Kaplan RM, Anderson JP. (1988) The general health policy model: Update and applications. Health Services Research 23, 203-235.

Kaplan RM, Bush JW, Berry CC (1976) Health status: Types of validity and the index of well-being. Health Services Research 11, 478-507.

Kazis L, Anderson JJ, Meenan RF (1989) Effect sizes for interpreting changes in health status. Medical Care 27 (Suppl), S110-127.

Kirshner B, Guyatt GH. (1985) A methodological framework for assessing health indices. Journal of Chronic Diseases 38, 27-36.

Koivukangas P, Koivukangas J, Ohinmaa A, Kivelä S-L, Krause K (1992) NHP - a method for measuring health-related quality of life in health services evaluation. Journal of Social Medicine 29, 229-235.

Liang MH, Fossel AH, Larson MG (1990) Comparisons of five health status instruments for orthopedic evaluation. Medical Care 28, 632-42.

Lovatt B. (1992) An overview of quality of life assessments and outcome measures. British Journal of Medical Economics 4, 1-7.

Lönnqvist J, Sintonen H, Syvälahti E et al. (1994b) Antidepressant efficacy and quality of life in depression: a double-blind study with moclobemide and fluoxetine. Acta Psychiatrica Scandinavica (forthcoming).

McDowell I, Newell C (1987) Measuring health: A guide to rating scales and questionnaires. Oxford University Press, New York, Oxford.

Patrick DL, Erickson P. (1993) Assessing health-related quality of life for clinical decision making. In Walker SR, Rosser RM. (eds.) Quality of life assessment. Key issues in the 1990s. Kluwer, Dordrecht, 11-63.

Pekurinen M, Vohlonen I, Sintonen H. (1991) Redefining incentives in primary health care: The Finnish demonstration project. In Lopez-Casasnovas G. (ed.) Incentives in health systems. Springer-Verlag, Heidelberg, 224-238.

Revicki DA (1989) Health related quality of life in the evaluation of medical therapy for chronic illness. The Journal of Family Practice 29: 377-

Rissanen P, Aro S, Slätis P, Sintonen H, Paavolainen P (1994) Health and quality of life before and after hip or knee replacement. Journal of Arthroplasty (forthcoming).

Rosser RM (1993) A health index and output measure. In Walker SR, Rosser RM. (Eds.) Quality of life assessment. Key issues in the 1990s. Kluwer, Dordrecht, 151-178.

Sintonen H. (1981a) An approach to economic evaluation of actions for health. A theoretic-methodological study in health economics with special reference to Finnish health policy. Official Statistics of Finland, Special Social Studies XXXII:74, Government Printing Centre, Helsinki.

Sintonen H. (1981b) An approach to measuring and valuing health states. Soc. Sci. 15C, 55-65.

Sintonen H. (ed.) EuroQol conference proceedings. Helsinki, October 1992. Discussion Paper No 2. Kuopio University Publications E. Social Sciences 8. Kuopio University Printing Office 1993.

Sintonen H, Lönnqvist J, Kiviruusu O. (1994) Cost-effectiveness/utility analysis of two drug regimens in the treatment of depression. National Centre for Health Program Evaluation, Working Paper 37, Melbourne.

Sintonen H, Pekurinen M. (1993) A fifteen dimensional measure of health-related quality of life (15D) and its applications. In Walker SR, Rosser RM. (Eds.) Quality of life assessment. Key issues in the 1990s. Kluwer, Dordrecht, 185-195, 467-470.

Sintonen H, Pekurinen M. (1989) A generic 15 dimensional measure of health-related quality of life (15D). Journal of Social Medicine 26, 85-96.

Slevin ML, Plant H, Lynch D, Drinkwater J, Gregory WM. (1988) Who should measure quality of life, the doctor or the patient? Br J Cancer 57: 109-112.

Statistics Finland (1993) Statistical yearbook of Finland 1993. Vol. 88. Printing Centre, Helsinki.

Stewart AL, Hayes RD, Ware JE. (1988) The MOS short-form general health survey. Reliability and validity in a patient population. Medical Care 26, 724-735.

Streiner DL, Norman GR (1989) Health measurement scales: A practical guide to their development and use. Oxford University Press, Oxford, New York, Tokyo.

Torrance GW, Boyle MH, Horwood SP. (1982) Application of multi-attribute utility theory to measure social preference for health states. Operations Research 30, 1043-1069.

APPENDIX 1

Quality of Life Questionnaire (New 15D)

Please read through all the alternative responses to each question before placing a cross (x) against the alternative which best describes your present status. Continue through all 15 questions in this manner, giving only one answer to each.

Question 1 ***Mobility***

1 ( ) I am able to walk normally (without difficulty) indoors, outdoors and on stairs.
2 ( ) I am able to walk without difficulty indoors, but outdoors and/or on stairs I have slight difficulties.
3 ( ) I am able to walk without help indoors (with or without an appliance), but outdoors and/or on stairs only with considerable difficulty or with help from others.
4 ( ) I am able to walk indoors only with help from others.
5 ( ) I am completely bed-ridden and unable to move about.

Question 2 ***Vision***

1 ( ) I see normally, ie I can read newspapers and TV text without difficulty (with or without glasses).
2 ( ) I can read papers and/or TV text with slight difficulty (with or without glasses).
3 ( ) I can read papers and/or TV text with considerable difficulty (with or without glasses).
4 ( ) I cannot read papers or TV text either with glasses or without, but I can see enough to walk about without guidance.
5 ( ) I cannot see enough to walk about without a guide, ie I am almost or completely blind.

Question 3 ***Hearing***

1 ( ) I can hear normally, ie normal speech (with or without a hearing aid).
2 ( ) I hear normal speech with a little difficulty.

3 ( ) I hear normal speech with considerable difficulty; in conversation I need voices to be louder than normal.
4 ( ) I hear even loud voices poorly; I am almost deaf.
5 ( ) I am completely deaf.

Question 4 Breathing

1 ( ) I am able to breathe normally, ie with no shortness of breath or other breathing difficulty.
2 ( ) I have shortness of breath during heavy work or sports, or when walking briskly on flat ground or slightly uphill.
3 ( ) I have shortness of breath when walking on flat ground at the same speed as others my age.
4 ( ) I get shortness of breath even after light activity, eg washing or dressing myself.
5 ( ) I have breathing difficulties almost all the time, even when resting.

Question 5 Sleeping

1 ( ) I am able to sleep normally, ie I have no problems with sleeping.
2 ( ) I have slight problems with sleeping, eg difficulty in falling asleep, or sometimes waking at night.
3 ( ) I have moderate problems with sleeping, eg disturbed sleep, or feeling I have not slept enough.
4 ( ) I have great problems with sleeping, eg having to use sleeping pills often or routinely, or usually waking at night and/or too early in the morning.
5 ( ) I suffer severe sleeplessness, eg sleep is almost impossible even with full use of sleeping pills, or staying awake most of the night.

Question 6 Eating

1 ( ) I am able to eat normally, ie with no help from others.
2 ( ) I am able to eat by myself with minor difficulty (eg slowly, clumsily, shakily, or with special appliances).
3 ( ) I need some help from another person in eating.
4 ( ) I am unable to eat by myself at all, so I must be fed by another person.
5 ( ) I am unable to eat at all, so I am fed either by tube or intravenously.

Question 7 Speech

1 ( ) I am able to speak normally, ie clearly, audibly and fluently.
2 ( ) I have slight speech difficulties, eg occasional fumbling for words, mumbling, or changes of pitch.

3 ( ) I can make myself understood, but my speech is eg disjointed, faltering, stuttering or stammering.
4 ( ) Most people have great difficulty understanding my speech.
5 ( ) I can only make myself understood by gestures.

Question 8 Elimination

1 ( ) My bladder and bowel work normally and without problems.
2 ( ) I have slight problems with my bladder and/or bowel function, eg difficulties with urination, or loose or hard bowels.
3 ( ) I have marked problems with my bladder and/or bowel function, eg occasional `accidents', or severe constipation or diarrhoea.
4 ( ) I have serious problems with my bladder and/or bowel function, eg routine `accidents', or need of catheterization or enemas.
5 ( ) I have no control over my bladder and/or bowel function.

Question 9 Usual activities

1 ( ) I am able to perform my usual activities (eg employment, studying, housework, free-time activities) without difficulty.
2 ( ) I am able to perform my usual activities slightly less effectively or with minor difficulty.
3 ( ) I am able to perform my usual activities much less effectively, with considerable difficulty, or not completely.
4 ( ) I can only manage a small proportion of my previously usual activities.
5 ( ) I am unable to manage any of my previously usual activities.

Question 10 Mental function

1 ( ) I am able to think clearly and logically, and my memory functions well
2 ( ) I have slight difficulties in thinking clearly and logically, or my memory sometimes fails me.
3 ( ) I have marked difficulties in thinking clearly and logically, or my memory is somewhat impaired.
4 ( ) I have great difficulties in thinking clearly and logically, or my memory is seriously impaired.
5 ( ) I am permanently confused and disoriented in place and time.

Question 11 Discomfort and symptoms

1 ( ) I have no physical discomfort or symptoms, eg pain, ache, nausea, itching etc.
2 ( ) I have mild physical discomfort or symptoms, eg pain, ache, nausea, itching etc.
3 ( ) I have marked physical discomfort or symptoms, eg pain, ache, nausea, itching etc.

4 ( ) I have severe physical discomfort or symptoms, eg pain, ache, nausea, itching etc.

5 ( ) I have unbearable physical discomfort or symptoms, eg pain, ache, nausea, itching etc.

Question 12 Depression

1 ( ) I do not feel at all sad, melancholic or depressed.

2 ( ) I feel slightly sad, melancholic or depressed.

3 ( ) I feel moderately sad, melancholic or depressed.

4 ( ) I feel very sad, melancholic or depressed.

5 ( ) I feel extremely sad, melancholic or depressed.

Question 13 Distress

1 ( ) I do not feel at all anxious, stressed or nervous.

2 ( ) I feel slightly anxious, stressed or nervous.

3 ( ) I feel moderately anxious, stressed or nervous.

4 ( ) I feel very anxious, stressed or nervous.

5 ( ) I feel extremely anxious, stressed or nervous.

Question 14 Vitality

1 ( ) I feel healthy and energetic.

2 ( ) I feel slightly weary, tired or feeble.

3 ( ) I feel moderately weary, tired or feeble.

4 ( ) I feel very weary, tired or feeble, almost exhausted.

5 ( ) I feel extremely weary, tired or feeble, totally exhausted.

Question 15 Sexual activity

1 ( ) My state of health has no adverse effect on my sexual activity.

2 ( ) My state of health has a slight effect on my sexual activity.

3 ( ) My state of health has a considerable effect on my sexual activity.

4 ( ) My state of health makes sexual activity almost impossible.

5 ( ) My state of health makes sexual activity impossible.