KM 29(1988)
pag 93 - 113

A measurement procedure for the evaluation of health states by the general public.

M.J. Verweij*, W.E. Saris*, G. Bonsel °,
B. van Hout °, J.D.F. Habbema °, P.J. van der Maas °, F.Th. de Charro °

Abstract

In this study the development of a measurement procedure for the evaluation of health states is discussed. This measurement procedure can be used to appraise the judgement of the general public about a large number of different health states. By this procedure quality-of-life scores are obtained on a fixed interval scale, which allows comparison across respondents. This procedure is shown to be reliable on individual level, and is also shown to be internally valid; furthermore, a relative high consensus is found between individuals.
In practice, up to 40 different health states can be evaluated in one interview session which takes about 25 minutes. This allows the possibility of collecting a large datapool concerning several different health states and calculating the relative gain in quality-of-life that can be obtained with specific medical therapies. Using these evaluations different health service programs can be compared not only on life-years gained but also on quality-of-life, and allocation decisions (where to spend how much money) can be improved.

*
Sociometric Research Foundation (SRF)
Blauwburgwal 20
1015 AT Amsterdam
Tel: 020-245641

°
Erasmus University Rotterdam
Postbus 1738
3000 DR Rotterdam
Tel: 010-4081111

Introduction

In medical decision-making research, various attempts have been made to quantify individual evaluations of different health states (e.g. Patrick et al.1973, Berg 1978, Sackett and Torrance 1978, Torrance 1986, Pauker and Kassirer 1987). Generally, a quality-of-life (QOL) score for a specific health state is obtained, in order to correct 'life-years' by the 'quality-of-life'. The result is called a quality-adjusted life-year (QALY; Pliskin et al. 1980). QALY-scores should in principle satisfy the properties of an utility-scale, and can be used to evaluate the outcomes of health service programs when medical decision making is involved, e.g. in cost-effectiveness analysis or other techniques for economic appraisal (Torrance 1986).
For correcting the life-years gained by their corresponding QOL-scores, a metric QOL-scale between 0 and 1 is required which allows the life-years gained to be multiplied by the QOL-scores. The purpose of this study was to find a procedure that meets the requirement stated above and that can be used on a large scale.

The previously proposed procedures turned out not to satisfy all theoretical and practical requirements (Sackett and Torrance 1978, Torrance 1986). With the existing procedures only a small number of health states can be evaluated, because they all require extensive instruction time and are too complicated for many respondents. This point will be returned to later.

In this study an effort is made to develop a reliable and valid measurement instrument for the evaluation of health states. The main requirement of this instrument is that it should be able to elicit social preferences for each of the health states. A consequence of submitting this particular judgement task to the general public is that medical terms have to be translated into layman's terms.

In this paper the accent lies on methodology. First a brief review of the existing evaluation procedures will be given. Then the design of the experimental study will be discussed and the results will be presented. After that follows a discussion and suggestions for further research are made.

Evaluation procedure.

In order to allow any application of QOL-scores in the above mentioned medical decision analyses, the measurement instrument should among other things meet the following requirements:

- the respondents must be capable of handling the task.
- the result should be a metric scale on (at least) interval level.
- there must be consensus among the respondents.
- the QOL-scores must be sufficiently informative i.e. accurate.

The most recent review of the literature on various procedures for evaluating health-states is given by Torrance (1986). The following procedures have been used: category rating, magnitude estimation, equivalence rating, standard gamble and time trade-off. Each of these procedures has its pros and cons.

Objections exist to category rating: due to the limited number of categories there is a lack of precision and reliability. Many different opinions, which otherwise could be distinguished, are pressed into one category on the scale (van Doorn et al. 1983). Also the required measurement on (at least) interval level is not obtained. Equivalence rating, standard gamble and time trade-off have as a major drawback that the required instructions and the task itself are complicated. For the general public, trade-offs are difficult to present (as will be shown in our pilot study). Furthermore, with regard to the standard gamble, a panel of 'experts' seems

necessary (see Torrance 1976), while its theoretical superiority is not supported by empirical evidence (Llewellyn-Thomas 1982). A consequence of the complexity of the latter three tasks is that in practice only a few health states can be judged and a relatively long interviewing time is needed.

Because magnitude estimation (m.e.), developed in psycho-physics by Stevens (1957), is both precise and relatively easy to understand and administer, as has been shown by Lodge (1982) and Saris (1987) for various topics, it is clear that m.e. has to be regarded as a good procedure for our purpose.
A physical equivalent of m.e. is 'line-production' (l.p.), where respondents express their opinion by means of the length of a line instead of a number. Use of the two procedures in combination provides the advantage of a test of the quality of the judgements of individual respondents. (Saris 1982)
Together m.e. and l.p. can be regarded as examples of the so-called 'comparison method' (Wegener 1982)

Our concrete goal was thus to obtain a metric QOL-scale with comparable judgement- scores between 0 and 1 for every health state for every individual.
This study intends to answer the following questions:

1. Are laymen capable of giving reliable evaluations of a large series of health states when using the comparison method (reliability, feasibility) ?
2. Are the respondents in agreement with each other (consensus) ?
3. Do the obtained QOL-scores give information that can be used in e.g. cost-effectiveness analysis (validity) ?

In order to answer these questions, two experiments have been carried out: a pilot study for testing the formulation of the questionnaire, and the main study for the calculation of the QOL-scores.

Measurement procedure

In this section we will discuss (a) the selection, construction and presentation of the health states descriptions (b) the response procedure and (c) the fieldwork.

(A) Health state descriptions

All relevant aspects (or dimensions) of a health situation should be incorporated into the health state description. In comparable research, different kinds of verbal descriptions of health states have been constructed and different combinations of dimensions have been made: either in an abstract two-dimensional way (Rosser and Kind 1982) or either in various multi-dimensional ways (Torrance 1986). In the multi-dimensional studies, the following dimensions are incorporated: (1) physical functioning or mobility, and (2) medical symptoms. Also, dimensions have been introduced that refer to (3) social relationships and (4) psychological well-being. Patient's socio and demographic information is nearly always presented (except when patients themselves were the jury) and sometimes the factor 'time spent in a particular state' is referred to (Culyer 1978, Torrance 1986).

All dimensions mentioned above are more or less necessary for the sake of completeness of the description, except the patient's demographic statistics. Since we restrict ourselves here to the question whether different health states can be evaluated (consistently), all aspects except the health situation itself are to be regarded as not relevant.

The relevance of the time aspect is not unequivocal. The evaluation of a health state indeed seems to depend on the past history (Sackett and Torrance 1978) and the future perspective, but this aspect is for this study regarded by us as one of many possible situational variables. In analogy with other researchers, the 'time spent' was however included in our first exploratory study.

It was decided to introduce two other aspects which are both doubtlessly relevant

when judging a health situation: (a) the intensity of the medical treatment, as indicated by the location of the patient (i.e. whether he is hospitalised or at home) and (b) the patient's subjective medical prognosis.

Treatment intensity was accounted for because the actual health states involved were derived from heart-transplant (HTX) patients. An important feature of these patients is the discrepancy between apparent health state and the need for intensive treatment.

Objective prognosis is less relevant, because every patient will interpret information about his prognosis according to his own references.

In the pilot study the following dimensions were used, which together formed one health state description or 'scenario':

- physical functioning
- medical symptoms
- social relationships
- psychological well-being
- time already spent in the health state
- location of the patient
- subjective medical prognosis

A large number of verbal descriptions of different health states were constructed, derived from the various HTX-phases. HTX-patients, both patients being investigated but not yet transplanted and transplant patients, have been under continuous investigation, so representative health pictures were available for each treatment stage.

In the main study, however, the social and the time dimension were excluded and the psychological and prognosis dimension were combined (see fieldwork section). In the Appendix the remaining dimensions are presented explicitly. Here, an example is given of a health situation as presented in the main study:

Somebody is:
- at home, under intensive out-patient control, taking powerful drugs
- rather quickly short of breath, at times in pain
- able to care for him/her self physically, though limited in walking around
- under stress, having trust in the future, but without looking too far ahead

Every effort has been taken to keep the descriptions as short and as comprehensible as possible, in order to lighten the task of the respondent.

(B) Response procedure

In general, when the comparison method is used, one asks for relative judgements, expressed against a given standard (Hamblin 1971, Lodge 1982). Recently it has been shown that variation in response, which is unrelated to the opinion expressed will occur, since respondents do not use the same scale. Presentation of two standards has been shown to prevent this otherwise uncontrollable variation (Saris 1987).

Consequently, we defined two standards. The upper one describes a health state in which somebody is perfectly healthy (the ideal health state):

Somebody is:
- at home
- no complaints or disorders
- able to do everything, both at home and outside
- no psychological complaints, and having trust in the future

The lower standard defines its opposite (the worst possible state):

Somebody is:
- in hospital, under 'intensive care'
- constantly in severe pain and out-of-breath
- restricted to bed, unable to do anything
- tense and depressed, living day by day and fearing the worst

These two standards were first introduced to the respondent, followed by two example questions. The wording of the l.p. question read as follows:

The IDEAL health state is given the following line:

The WORST POSSIBLE health state is given this line:

–

Please, draw your line for the following health state:

Somebody is:

(here a specific health state description is presented)

It is clear that for magnitude estimation an analogous question can be formulated, where as upper standard the number 1000, and as lower standard the number 1, are presented.

C) Fieldwork

In this methodological research two subsequent experimental studies were carried out (the previously mentioned pilot study and the main study, five months later) in order to test the QOL-evaluation procedure proposed above. We made use of computer assisted interviewing (Jacklin 1984, Saris and De Pijper 1986). Considerable improvement in measurement is obtained this way using (1) continuous scales (van Doorn et al. 1983), (2) different response modalities for correction of measurement error (Saris 1982) and (3) facilities for automatic routing, random presentation, avoidance of missing data, etc.

In our studies the SRF-panel was involved, which consists 44 Dutch households based on a random population sample. These households cannot be seen as representing any population. However, there is enough variation in background for performing tests of measurement procedures. This panel has been in existence for one year now and the respondents are used to answering interviews by computer on a monthly basis (mostly for methodological research; often including the use of l.p. and m.e.). As a reward the panel members are given a home-computer in loan and occasionally some new software.

Here we will present some of the background variables, in order to show the

variation in sex (Table 1), age (Table 2), income (Table 3) and education (Table 4).

Sex		%
Male	35	48.6
Female	37	51.4

Table 1. Sex (N=72)

Age		%
< 20 years	15	21.0
20 - 30	13	18.0
30 - 40	15	21.0
40 - 50	12	16.6
50 - 60	9	12.5
60 - 70	7	9.7
70 +	1	1.4

Table 2. Age (N=72)

Income, if employed		%
<1000	5	15.2
1000-1500	4	12.1
1500-2000	4	12.1
2000-2500	8	24.2
2500-3000	5	15.2
3000-3500	3	9.1
>3500	4	12.1

Table 3. Income, if respondent has a job (N=33)

Education		%
primary school	4	5.6
lower vocational	14	19.4
secondary school	16	22.2
middle vocational	5	7.0
college	13	18.0
higher vocational	15	21.0
university	5	7.0

Table 4. Education (N=72)

With the pilot study we mainly aimed at getting an answer to the following questions: what is the best way of presenting the questions?; is the content of the descriptions understandable for laymen?; does the internal order of the dimensions make any difference to the given evaluation score?; are all dimensions necessary? and how much time takes an average interview session?

In the pilot study (N=72) only theoretical health states were presented, based on seven dimensions (as stated before). After introducing the topic and the standards, 21 different health states were presented twice, in random order; the first round using l.p. as the response modality, the second round using m.e. In the pilot study we tested the 21 theoretical health descriptions in a so-called factorial design. In this way the health states could be formulated independently of each other, whereby health states that are too unrealistic were avoided. It is obvious that in

reality, when going from better to worse health states, all dimensions will be changing into the same direction to some degree (see Rosser and Kind 1982).

The factorial design of theoretical descriptions could thus prevent most of this 'natural' multi-collinearity, so that a multiple regression analysis on the average QOL-scores could be performed, with every single dimension as predictor. Together the predictors could explain more than 88 percent of the total variance, which was a very encouraging result, as it meant that the procedure measured what we set out to measure. 'Social relationships' explained only 2 percent of the total variance. We suppose that this aspect is considered relatively unimportant, because of its situational character. It seems plausible that the social relations of somebody else will not be taken into account when judging his health state. So, for the main study, we decided to exclude this dimension.

Thinking again about the role of the 'time spent' dimension, we concluded that in reality the time factor will be inseparably associated with psychological aspects; the psychological well-being of a patient will be determined by both his past and his prognosis. A time factor will thus already be accounted for in the 'subjective prognosis'. Double counts would seriously invalidate the resulting QOL-scores and consequently any further medical decision analysis. So, on theoretical grounds, we decided to exclude this dimension as well.

In the pilot study we also tried out three time trade-off and three money trade-off tasks as intervening judgement-tasks between the l.p. and m.e. series. These trade-off tasks turned out to be too complex for the respondents, and resulted in non-interpretable data. About half of the respondents gave ostensibly nonsense-responses and 20% refused to answer. The small number of answers which could be interpreted were unreliable, because nearly identical responses were given on different trade-offs, while different responses were given on the same trade-off.

The trade-off tasks also prolonged the interview session unacceptably; they caused an extension of the average interviewing time of 20 minutes.

The main study was meant to test the corrected design. The trade-off tasks were thus deleted and only four dimensions (see Appendix) were presented. The formulation of the health states was again shortened and clarified wherever this was possible. This time 26 different health descriptions were presented and evaluated in lines; 11 of them were later repeated in numbers.
Consequently, the average interviewing time could be reduced to 25 minutes absolute.

In the following part of the paper the main results of the statistical analyses for the main study will be reported.

Results

(A) Reliability of the line/number responses

For every respondent we have at our disposal two independent measurements of the evaluation of the health states (one given in lines and one given in numbers). These judgements are expressed on the same scale, namely between the two extreme health states which are presented as standards. The relation between the line-scores and the number-scores can be studied for every respondent separately. If there is no error, then the judgement scores obtained via one method (e.g. lines) should correlate perfectly with the judgement scores obtained via the other method.
In this type of analysis, which obviously is only possible with repeated measurement of the same questions, one can find for many topics a very high correlation on individual level between judgements gathered by the comparison method, providing that the respondent is sufficiently acquainted with the topic under investigation. In Table 5 an overview is given of similar research findings on other topics.

Topic	Median Correlation	Source
Positions of political parties (left / right)	.95	van Doorn, Maas and Saris (1987)
Evaluation of work load	.95	Zijlstra and van Doorn (1987)
Satisfaction with income	.95	van Praag and van Doorn (1987)
Evaluation of family size	.92	van Doorn (1985)
Evaluation of time spending	.86	van Doorn (1985)
Evaluation of housing situation	.84	van Doorn (1985)
Evaluation of relationship-patterns	.80	van Doorn (1985)
Evalution of life-satisfaction	.79	van Doorn (1985)
Evaluation of work-aspects	.75	Saris and Prins (1985)
Importance of household activities	.56	Henstra, van Doorn (1985)

Table 5: The median correlation of evalution scores expressed in lines and numbers for a number of other topics.

A median correlation between the line and number scores of 0.85 and higher is a quite normal result. This is a necessary criterion, in order to be able to conclude that the respondents can handle the required task.

In the pilot study the median of correlations was quite low, only 0.67. At that stage we supposed that the health descriptions were still too complex, and therefore cost the respondents too much time to read and comprehend. Therefore, the design for the main study was improved by (1) the exclusion of the social and time dimension (as stated above) and (2) the shortening, clarification and simplification of the formulation of the remaining dimensions.

For our topic the results for the main study are given in Table 6.

mean of correlations	.89
median of correlations	.92
interquartile-distance	.89 - .95

Table 6. Mean correlation of the line versus the number judgements of the health states. The median correlation and the interquartile-distance (q_1-q_3) are also given.

In this table one can see that the median correlation between the line and number scores is considerably above our criterion. That the length and the formulation of

the health states indeed caused confusion (resulting in a low intra-reliability) may be clear from the gain in the median correlation obtained (from 0.67 to 0.92).
With the improved formulation, nearly every respondent's individual scales correlate highly with each other, which means that these scales are approximately the same. Because the respondent's judgements are consistent, we conclude that laymen are able to do the job.
It was also checked whether differences existed in individual reliability associated with the background variables: age, education and a third variable, that measures the amount of experience one has in the field of health care (subjectively expressed). No significant association was found.

(B) Consensus

The next question is whether one QOL-scale exists for everybody, or whether different subgroups - i.e. where the opinion deviates explicitly - need to be distinguished. Further analysis with aggregated QOL-scores will only then be justified when the consensus is proven to be high. The existence of subgroups is investigated by (1) calculating the amount of consensus (2) checking on 'out-liers' as possible source of deviation (3) splitting up according to background characteristics.

In order to analyse in how far respondents are in agreement, one can investigate to what extent their individual QOL-scales correspond with the 'group scale', which is the QOL-scale aggregated across respondents. For this association to be high, both a high intra-reliability and a high consensus are required. In Table 7 shows that high consensus does not always exist.

Topic	Correlation	Source
Work load	0.95	Zijlstra, van Doorn (1987)
Work experience	0.86	Saris e.a. (1977)
Position of political parties	0.75	van Doorn, Maas, Saris (1987)
Quality of work	0.66	Saris e.a. (1977)
Risk of political protest actions	0.53	Saris e.a. (1977)
Effectivity of political protest actions	0.22	Saris e.a. (1977)

Table 7. Consensus in judgement based on the correlation between individual and and group scales.

We found that in the main study the correlation between the individual line scores and the group scale was 0.86. We may thus conclude that laymen agree to a large extent about the degree of seriousness of different health states, as presented in this way. Also, no significant association was found between the three background characteristics and the 'individual scale/group scale' correlation.

Although these results are very good on aggregate level, they do not exclude the possibility of individual variation in scores. The observed standard deviations of the evaluated items were around 0.16 on the 0 - 1 scale (see Table 8). This might be due (1) to respondents with an intra-reliability lower than 0.85 or (2) to respondents with a certain difference in opinion or (3) to variation in response behaviour. When excluding the scores of the 9 respondents with an intra-reliability lower than 0.85, the standard deviation decreases indeed, but not much (around 0.01). After also excluding the scores of the 4 respondents with a small difference in opinion and the 4 respondents with variation in response behaviour, the s.d. decreases again but not much (around 0.02). We thus concluded that the 'deviators' were not deviating to such an extent that their exclusion leads to a large reduction of the variability. As these deviations are hardly larger than the variability in the individual answers of non-deviating respondents, they can therefore be ignored.

There are also no distinctive sub-groups, in which a difference in opinion could be distinguised. In earlier studies it was sometimes shown that response scores concerning health states can be related to characteristics of respondents (e.g. their age: Sackett and Torrance 1978). We thus wanted to test in more detail whether or not the data would indeed show similar differences in QOL-scores for different subgroups of subjects.

The following subgroups were therefore created:

- low educated respondents (N=35) versus high educated respondents (N=30)
- respondents younger than 40 years (N=36) versus older than 40 years (N=29)
- respondents who don't consider themselves experienced in the field of health care (N=51) versus respondents who do consider themselves experienced (N=15).

Next we calculated the 'subgroup QOL-scale' for every subgroup.

Since we are claiming that the QOL-scale should be the same for every sub-group, the coefficient of identity (Zegers and Ten Berge 1985) - reflecting the degree to which two scales are identical (=same mean, same dispersion, same distribution form) - should be close to unity when no differences exist. This coefficient was found to be 0.993 between the subgroup scales of the two education levels, 0.997 between the two age groups, and 0.998 between the 'real laymen' and the 'semi-professionals'. This means that even with this very strict measure no differences in QOL-judgement between subjects could be found.

(C) Validity

Finally, the obtained QOL-scales will be evaluated. First, the possibility of order-effects within the health description will be examined. Then the group-scale obtained by drawing lines will be compared with the group-scale obtained by assingning numbers. Again the total amount of explained variance is reported and the width of the confidence intervals is considered.

It can be hypothesized that a different order of dimensions in the presentation of the health description can lead to a different judgement score. In the pilot study some evidence in support of an internal order-effect was found. However, it can be expected that this phenomenon will only occur when the list of dimensions is long. We hypothesized that with a short, four-dimensional health description, this order-effect will not occur because the health description is then regarded and judged as a whole.
In order to test this hypothesis, two health descriptions were presented twice, but with varying internal ordering. No difference in the resulting mean judgement scores was found. We therefore have no reason to believe that possible order effects should be taken into account, as long as the health description is short and comprehensible.

We have obtained the mean judgements about different health states expressed both in lines and - independently - in numbers. The coefficient of identity between both group scales is a measure of the meaningfulness or the internal validity of such a group scale. This coefficient was equal to 0.98 and the aggregated groups scales are thus completely exchangeable.
Another measure of validity is the amount of variance that the four dimensions are together able to predict of the QOL-scale. When performing this regression analysis across persons, a percentage of 83% explained variance was found. This indicates that almost all variation can be deduced from what was put into the health descriptions and does not come from other unknown sources.

In Table 8 the QOL-scale is presented, based on the line judgements for 26 different health states. The different health states are indicated by a 4 digit number, each digit representing a category on a dimension. The dimensions and categories are ordered as in the Appendix. The scores are given on a 0 - 1 scale and the corresponding 95% confidence intervals are given.

Health State	Mean QOL-score	Median	s.d.	s.e.	95% Conf. Interval
1111.	.97	1.00	.06	.007	.95 - .98
3111.	.61	.63	.18	.023	.56 - .65
2222	.51	.50	.18	.021	.46 - .54
2232.	.48	.48	.19	.024	.43 - .52
1331.	.47	.44	.19	.023	.42 - .52
2322.	.44	.44	.17	.021	.39 - .49
4222.	.44	.43	.19	.023	.43 - .52
3114.	.41	.42	.19	.023	.36 - .45
1323.	.40	.39	.17	.021	.36 - .44
2224.	.37	.39	.16	.020	.33 - .41
3321.	.37	.36	.16	.020	.33 - .41
4432.	.37	.37	.18	.022	.32 - .41
3323.	.34	.32	.18	.022	.30 - .38
3333.	.33	.32	.16	.019	.30 - .37
5232.	.31	.29	.19	.023	.27 - .36
2442.	.30	.30	.17	.021	.26 - .34
3343.	.29	.29	.15	.018	.25 - .32
2434.	.29	.29	.16	.019	.25 - .33
4434	.27	.26	.15	.018	.23 - .31
4444.	.22	.21	.14	.017	.18 - .25
4553.	.20	.18	.13	.016	.17 - .23
5443.	.17	.13	.12	.015	.14 - .19
4454.	.16	.16	.11	.013	.14 - .19
5543.	.15	.12	.12	.015	.12 - .18
4545.	.13	.09	.10	.013	.10 - .15
5555.	.03	.00	.02	.002	.03 - .04

Table 8. Detailed data description of the QOL-scores of various health states, expressed on a 0-1 scale, for the whole sample (N=66).

In Table 8 the health states are ordered according to their QOL-score: from better to worse. The whole response-continuum is spanned, except for a gap between the first two states. An ordinary influenza description would probably fit in this gap. Most of the confidence-intervals show overlap. It is quite simple to calculate how large a representative sample for the population should be in order to decrease this overlap to a minimum. Most of the overlap is avoided if the 95% confidence interval has as its limits: mean $\pm$ 0.01. In that case N should be 1110.

Discussion

In this study it has been shown how a reliable and internally valid measurement instrument for the evaluation of health states is obtained by using the magnitude estimation procedure. All respondents were able to handle the evaluation task without difficulty. Furthermore, consensus among respondents was found.
The resulting QOL-scale is a metric scale between 0 and 1 and the scores are comparable over respondents.
Nothing can be claimed about the external validity. An examination of the relationship of the QOL-scale with the scales obtained by other series of health states, remains a point for further research.
This measurement procedure can be used for the general public on a large scale. We have shown that up to 40 different health states can be evaluated in one interview session, taking less than 25 minutes of interviewing time.
This gives the possibility of collecting a large datapool concerning several different health states. We recommend that at least 8 health states should always be added as anchor-points. This is advised in order to check if the anchor-points keep the same QOL-score when imbedded in series of descriptions of other diseases. This can guarantee comparability of the scales across studies.
Likewise several medical health programs, for example heart transplantation, kidney transplantation and liver transplantation, can be compared with regard to their costs and effective outcome, not only taking the number of life-years gained into account, but also the relative gain in quality-of-life.

References

Berg R.L. (1973)
Establishing the values of various conditions of life for a health status index. In: Berg (ed) Health status indexes.

Culyer A.J. (1978)
Measuring Health. Lessons for Ontario. University of Toronto Press.

Doorn L. van, Saris W.E., Lodge M. (1983)
Discrete or continuous measurement: what difference does it make ? Kwantitatieve Methoden, 10.

Doorn L. van (1985)
Eindverslag van het onderzoek naar de kwaliteit van de tevredenheidsvragen van het Leefsituatieonderzoek. Amsterdam. Sociometric Research Foundation.

Doorn L. van, Maas C.F. and Saris W.E. (1987)
Different procedures to measure left-right positions of parties and voters. Acta Politica (forthcoming)

Hamblin R.L.(1971)
Social attitudes: magnitude measurement and theory. In H.M. Blalock (ed) Measurement in the social sciences. London, MacMillan Press.

Henstra C. and van Doorn L. (1985)
Evaluatie van huishoudelijke productie. Research Memorandum 150785. Sociometric Research Foundation. Amsterdam.

Jackling P. (1984)
Computer assisted questionnaire design: the real breaktrough. In ESOMAR: Are interviewers obsolete ? Drastic changes in data collection and data presentation. European society for opinion and marketing research (ESOMAR). Amsterdam.

Llewellyn-Thomas H. and Sutherland H.J. et al. (1982)
The measurement of patients' values in medicine. Medical decision making, 2, 4.

Lodge M. (1982)
Magnitude scaling: Quantitative measurement of opinions. Beverly Hills: Sage.

Pauker S.G. and Kassirer J.P. (1987)
Decision analysis. In: The New England Journal of Medicine, 316.

Pliskin J.S., Shepard D.S. et al. (1980)
Utility functions for life years and health status. Oper. Research, 28.

Praag B.M.S. van, Doorn L. van (1987)
Measurement procedures for income satisfaction compared. In: W.E. Saris (ed). Variation in response functions: a source of measurement error in attitude research. Amsterdam. Sociometric Research Foundation.

Rosser R.M. and Kind P. (1982)
A scale of valuations of states of illness: is there a social consensus ? International Journal of Epidemilogy, 7, 4.

Sackett D.L. and Torrance, G.W. (1978)
The utility of different health states as perceived by the general public. J. Chronic Disease, 31.

Saris W.E., Bruinsma C., Schoots W., and Vermeulen C. (1977)
The use of magnitude estimation in large scale survey research. Mens en Maatschappij, 52.

Saris W.E.(1982)
Different Questions, Different variables. In: C. Fornell (ed) Second generation of multivariate analysis. New York. Praeger Publishers.

Saris W.E. and Prins P. (1986)
Evaluation of work aspects: a comparison between the USA and Germany. Amsterdam. Sociometric Research Foundation.

Saris W.E. and W.M. de Pijper (1986)
Computer assisted interviewing using home computers. European Research, 14, 3.

Saris W.E. (1987)
Variation in response functions: a source of measurement error in attitude research. Amsterdam. Sociometric Research Foundation.

Stevens S.S. (1957)
On the psychophysical law. In: Psychophysical review, 64.

Torrance G.W. (1976)
Social preferences for health health states: an empirical study of three measurement techniques. In: Socio-Economic Planning Sciences, 10.

Torrance G.W. (1986)
Measurement of health state utilities for economic appraisal: a review. J. of Health Economics, 5.

Wegener (1982)
Social attitudes and psychophysical measurement. Hillsdale: Earlbaum.

Wolfson A. (1974)
A health index for Ontario. Toronto, Ministry of Treasury and Intergovernmental Affairs.

Zegers F.E. and ten Berge J.M. (1985)
A family of association coefficients for metric scales. Psychometrika, 50,1.

Zijlstra F. and Doorn. L. van (1987)
Variations in response functions complicates the evaluation of scales; in W.E. Saris (ed). Variation in response functions: a source of measurement error in attitude research. Amsterdam. Sociometric Research Foundation.

Ontvangen: 02-04-1987
Geaccepteerd: 14-06-1988

Appendix

The translated wording of the health dimensions as used in the main study.

A. Location (intensity of treatment)

1- just at home
2- at home, but having regular medical check-ups
3- at home, under intensive out-patient control, taking powerful drugs
4- temporarily hospitalized
5- in hospital, under 'intensive care'

B. Medical Symptoms

1- no complaints or disorders
2- some small complaints, but not in pain
3- rather quickly short of breath, at times in pain
4- at the smallest effort short of breath and tired, in moderate pain
5- constantly in severe pain and out-of-breath

C. Physical Functioning

1- able to do everything, both at home and outside
2- able to do everything at home, but restricted outdoors (e.g. unable to cycle or do the shopping)
3- able to care for him/herself physically, though limited in walking around
4- difficulties with getting in/out of bed, needing help with self-care
5- restricted to bed, unable to do anything

D. Psychological well-being + subjective medical prognosis

1- no psychological complaints, and having trust in the future
2- feeling down now and then, but having trust in the future
3- under stress, having trust in the future, but without looking too far ahead
4- depressed, moderate trust in the future, does not look far ahead
5- tense and depressed, living day by day, and fearing the worst