An Experimental Test of a Theoretical Foundation for Rating-scale Valuations

HAN BLEICHRODT, PhD, MAGNUS JOHANNESSON, PhD

A major advantage of using a rating scale in health-utility measurement is its practical applicability: the method is relatively easy to understand, and various health states can be assessed simultaneously. However, a theoretical foundation for rating-scale valuations has not been established. The primary aim of this paper is to present a theoretical foundation for rating-scale valuations based on the theory of measurable value functions and to provide a consistency test to see whether rating-scale valuations do indeed elicit a measurable value function. If rating-scale valuations elicit a measurable value function, then Dyer and Sarin have shown how they are related to von Neumann-Morgenstern (vNM) utilities. The appropriate technique to measure vNM utilities is the standard gamble. Torrance has suggested that rating-scale valuations and standard-gamble valuations are related by a power function. A secondary aim of this paper is to examine the relationship between rating-scale valuations and standard-gamble valuations hypothesized by Torrance. An experiment was designed to test consistency of rating-scale valuations and the relationship between rating-scale valuations and standard-gamble valuations. The experiment tested whether rating-scale valuations are independent of the context in which they are elicited, as they should be if they elicit points on a measurable value function. 80 Swedish and 92 Dutch respondents participated in the experiment. The results showed that rating-scale valuations depend on the number of preferred alternatives in the task and thus violate a basic property of measurable value functions. The estimation of the power function did not result in stable results: parameter estimates varied, in some cases there was indication of misspecification, and in most cases there was indication of heteroskedastic errors. The implications of these findings for the common use of rating-scale valuations in cost-utility analysis are serious: the dependency of the rating-scale valuations on the other health states included in the task casts serious doubts on the validity of the rating-scale method. *Key words:* QALYs; rating scale; cost-utility analysis; medical decision making. **(Med Decis Making 1997;17:208-216))**

Quality-adjusted life years *(QALYs)* are by far the most commonly used outcome measure in cost-utility analyses of health care programs. The major advantage of QALYs is that they simultaneously take into account quality of life and quantity of life. To calculate QALYs, life years are adjusted for quality of life by multiplying life years by a weight that reflects the utility of the health state in which these years are spent. Three main methods exist to estimate these quality weights: the rating-scale (RS), the time-tradeoff (TTO), and the standard gamble (SG) (for a description of these methods see, for example, Drummond et al.[1]).* To date, no consensus exists in the literature on the question of which method to use. Justifications for the different meth- ods can, for example, be found in Torrance and Feeny,[3] Broome,[4] and Richardson.[5]

Unfortunately, empirical research has shown that the three methods, given common scaling, lead to different valuations.[6-8] The typical pattern is that the valuations elicited by the standard gamble are higher than the valuations elicited by the time trade-off, which in turn are higher than the valuations elicited by the rating scale. The differences between the various methods are in general statistically significant. This discrepancy raises the question of which method to use. To answer this question, at least three considerations deserve attention. First, which method is easiest to use? Second, what exactly is measured by the various methods? In particular, do the methods lead to valuations that capture what we are interested in in cost-utility analyses? Third, does a stable relationship exist between the valuations elicited by the various meth-

Received September 27, 1995, from the Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University, Rotterdam, The Netherlands (HB); and the Centre for Health Economics, Stockholm School of Economics, Stockholm, Sweden (MJ). Revision accepted for publication March 25, 1996. Supported by the National Corporation of Swedish Pharmacies and by Merck, Sharp and Dohme.

Address correspondence and reprint requests to Dr. Bleichrodt: iMTA, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. e-mail: (bleichrodt@econ.bmg.eur.nl).

*There exist other approaches to estimate quality weights for health states, for example the person-tradeoff technique advocated by Nord.[2]

ods? With respect to the first question, the common finding in practical applications is that the rating scale is easiest to use, in the sense that the method is simple and cheap to develop and quick and efficient to use. With respect to the second question, various authors have advocated the standard gamble on the grounds that it has a well-established foundation in von Neumann-Morgenstern (vNM) expected utility theory and therefore what is measured are vNM expected utilities. The rating-scale method lacks a foundation in axiomatic decision theory. Finally, with respect to the third question, if a stable relationship between the methods were to exist, the question of which method to use would become less urgent, because in that case one could easily convert the valuations elicited by the various methods into each other. We return to the existence of a stable relationship between the elicitation methods in the sequel of this paper.

Given the relative ease with which the rating scale can be applied, it can be argued that if one could succeed in providing the rating scale with a satisfactory foundation in decision theory, it would be the preferred method. The aim of the present study was to examine a theoretical foundation of the rating scale based on the interpretation that is commonly given to rating-scale valuations. We argue that in the common interpretation given to rating-scale scores, the rating-scale method should elicit a so-called "measurable value function." We then present a consistency test to examine whether the rating-scale method indeed elicits a measurable value function.

Theoretical Background

A common view among researchers involved in measuring health-state utilities is that the difference between the rating scale on the one hand and the standard gamble on the other is that the standard-'gamble question is framed in terms of risk whereas the rating-scale question is framed in terms of certainty. The difference between valuations elicited by the standard gamble and by the rating scale is explained by attitude towards risk. According to this explanation, the rating scale measures preferences under certainty and the method elicits an underlying value function for health. For example, Torrance et al.[9] observe with respect to the rating-scale method (to which they refer as "category scaling method"): "Because uncertainty is not used in the category scaling method, it measures a value function $v_i^*(x_i^*)$ as opposed to a utility function $u_i^*(x_i^*)$."

Moreover, this value function is supposed to measure differences in preferences. As Torrance*' explains the rating scale valuations: " . . . the intervals between the placements correspond to the differences in preference as perceived by the subject. [p. 18]" Thus, the value function elicited by the rating scale not only orders alternatives according to preferences, it also orders differences in preferences between alternatives.

So, ideally, the rating scale measures a value function that reflects strength of preference. Such a value function is known in the literature as a "measurable value function," to distinguish it from the more common value function that represents preferences under certainty, but does not reflect strength of preference.[11] Measurable value functions have been axiomatized by Krantz et a1.[12] and studied in detail by Dyer and Sarin.[13,14] It is worthwhile to consider briefly the theory behind such measurable value functions. (For more details the reader is referred to the papers by Dyer and Sarin.) Consider a preference relation $\geq^*$ over pairs of alternatives. We interpret $pq \geq^* rs$ to mean that the strength of preference for health state p over health state q is at least as great as the strength of preference of health state r over health state s. If the axioms formulated by Krantz et al. hold, the preference relation $\geq^*$ can be represented by a measurable value function v. That is, $pq \geq^* rs$ if and only if

$$v(p) - v(q) \geq v(r) - v(s) \qquad (1)$$

Two properties of v are worth noting. First, v is unique up to positive linear transformations. That is, if v' also satisfies equation 1, then there exist real numbers $\alpha > 0$ and β such that $v'(q) = \alpha v(q) + \beta$ for all health states q. Thus, if we scale both $v'(q)$ and $v(q)$ such that, for example, v' (full health1 = v (full health) = 1 and v' (immediate death) = v (immediate death) = 0, then it follows straightforwardly that v' $(q) = v(q)$. So, given common scaling, all measurable value functions should lead to the same valuations for all health states q. A second property of v is that it is also a value function in the more common sense of the term. Define a preference relation $\geq$, meaning "at least as preferred as," from $\geq^*$ by requiring that $p \geq q$ if and only if $pr \geq^* qr$ for all health states p, q, and r. Thus, $p \geq q$ if and only if $v(p) - v(r) \geq v(q) - v(r)$, or $p \geq q$ if and only if $v(p) \geq v(q)$ and thus v represents $\geq$ as well. So measurable value functions represent preferences under certainty.

Dyer and Sarin[14] describe three approaches to assess a measurable value function. The first approach is to simply set v equal to a vNM utility function. However, the idea that vNM utilities measure strength of preference has been rejected by such influential authors as Arrow," Ellsberg,[16] and Fishburn,[17] among others. Sarin[18] has shown that vNM utilities measure strength of preference only if Savage's sure-thing principle is strengthened. A second approach is to define $pr \geq^* qs$ in terms of another attribute. For example, one can conclude that pr $\geq$

*qs if an individual is willing to pay at least as much for a health improvement from r to p as for a health improvement from s to q. This approach requires that preference differences not be affected by the attribute by which strength of preference is assessed (Dyer and Sarin refer to this condition as "difference independence").

A third approach proposed by Dyer and Sarin is to take strength of preference as a primitive concept and to elicit strength of preference from introspection. This approach does not use choices to elicit strength of preference, as do the two previous approaches, but requires an individual to indicate directly his or her strength of preference. This is the approach used in the rating-scale method: individuals are asked to put health states on a line where differences between placements reflect strengths of preferences and choices are not used to elicit preference differences.

As we observed above, a measurable value function is not necessarily equivalent to a vNM utility function. However, as has been argued by Dyer and Sarin,[14] there exists a straightforward relationship between the two. A vNM utility function is the product of at least two factors, which cannot be separated: 1) strength of preference and 2) attitude towards risk. The usual Arrow-Pratt measure of risk attitude reflects both these factors. Dyer and Sarin define an alternative measure of risk attitude that reflects only risk attitude and not strength of preference and to which they refer as "relative risk attitude." The idea behind the approach of Dyer and Sarin is. first to assess a measurable value function v, which reflects strength of preference. Then relative risk attitude is incorporated in this measurable value function to arrive at the vNM utility function u. Thus, a vNM utility function is a function of a measurable value function $\{u = u[v(x)]\}$ where the shape of this function reflects an individual's relative risk attitude. If $u[v(x)]$ is concave, for example, the individual is defined to be relatively risk-averse.

The theoretical relationship between measurable value functions and vNM utility functions corresponds to a belief in health-state utility measurement that a stable relationship exists between rating-scale valuations and standard-gamble valuations.[6] In particular, Torrance has suggested that standard-gamble valuations and rating-scale valuations are related by the following formula:

$$1 - SG = (1 - RS)^{\beta} \qquad (2)$$

where RS stands for "rating-scale valuation" and SG stands for "standard-gamble valuation." In this relationship, β could then be interpreted as the coefficient reflecting relative risk attitude. Torrance found a stable relationship at the group level, which gave a good fit.

Several conclusions can be drawn from the above exposition. First, the interpretation generally attached to rating-scale valuations suggests that this method attempts to elicit a measurable value function. Second, given that vNM utility functions and measurable value functions are equivalent only under certain, fairly restrictive, assumptions, in general, standard-gamble valuations will differ from rating-scale valuations. Third, there exists a relationship between vNM utility functions and measurable value functions that may take the form suggested by Torrance.

Hypotheses

As noted in the introduction, the main aim of this study was to examine the theoretical foundation of the rating scale. The previous section suggests a foundation for rating-scale valuations: they may elicit a measurable value function. If the rating-scale valuations do indeed elicit a measurable value function, then we have seen in the above section that, given common scaling, identical health state valuations should result from different rating-scale tasks. That is, if we set $v(\text{full health}) = 1$ and $v(\text{immediate death}) = 0$, then $v(q)$ should be the same in different contexts.

The null hypothesis we test in the experiment reported below is that the rating scale elicits a measurable value function and that the $v(q)$ will be the same in different contexts. However, a danger in rating-scale exercises is that context effects affect the results; valuations may, for example, depend on the other health states that are included. Respondents in a rating-scale task may have a tendency to spread health states over the whole scale. Thus, if the other health states that are included are relatively attractive, $v(q)$ will be relatively low, whereas if the other health states that are included are relatively unattractive, $v(q)$ will be relatively high. As Loomes et al.[19] have observed, this hypothesis corresponds to Parducci's range-frequency model.

The above consistency test is the main empirical test we carried out in this study. Even if we obtain identical results in different contexts, we cannot yet conclude that the rating-scale method does indeed elicit a measurable value function. However, if we find different valuations in different tasks, then we can conclude that the rating-scale valuations violate a basic property of measurable value functions and that we should seriously worry about what is measured by the rating scale.

A secondary hypothesis that we test is whether a stable relationship of the form hypothesized by Torrance exists between rating-scale valuations and standard-gamble valuations. Notice that this research question makes sense only if the rating-scale valuations do indeed elicit a measurable value function. It should be emphasized that there are no a priori theoretical reasons why the relationship between a measurable value function and a vNM utility function should take the form given in equation 2. However, this form has the important advantage of taking into account the scaling of the data: if the standard-gamble valuation is equal to 1, then the rating-scale valuation is also equal to 1, and if the standard-gamble valuation is equal to 0, then the rating-scale valuation is also equal to 0. If we do not obtain a stable relationship, this could be due to several factors: for example, the rating scale does not elicit a measurable value function even though it passes the consistency test described above, the valuations elicited by rating scale and by the standard gamble are distorted by biases, or the relationship does not take the form hypothesized by equation 2. Thus, if we do not obtain stable results in the test for this secondary hypothesis, this indicates only that something is wrong, but we are not able to say what exactly is wrong.

Methods

RESPONDENTS AND HEALTH STATES

The respondents were 80 students at the Stockholm School of Economics and 92 students at Erasmus University Rotterdam. All were undergraduates recruited from courses in economics, statistics, and health policy. They were paid approximately $15 for their participation. The experiment was carried out in 17 sessions lasting approximately one hour with, on average, ten respondents per session. The procedure in each session was to explain a specific task to respondents, obtain their responses to this task, and then to move on to the next task. A “master” version of the experiment was designed in English. This “master” version was subsequently translated into Swedish and Dutch. Before drafting the final version, we tested the questionnaire extensively both in Stockholm and in Rotterdam using faculty staff members as respondents.

We include8 eight health states in the questionnaires. The health states were taken from the Maastricht Utility Measurement Questionnaire,” a slightly adapted version of the McRheum, a McMaster health utility measurement designed specifically for rheumatoid arthritis.[21] The selected health states correspond to commonly, occurring types of back pain and rheumatism. The health states consist of four dimensions. The health classification system is shown in table 1. The health states were described on a set of cards, which were handed out to respondents at the beginning of each session. The health states (A-H) used in the experiment are described in table 2.

Table 1 • The Multi-attribute Health-status-classification System Used in the Experiment

General daily activities
Able to perform all tasks at home and/or work without problems
Able to perform all tasks at home and/or work, albeit with some difficulties
Not able to perform some tasks at home and/or at work
Not able to perform many tasks-at home and/or at work
Not able to perform any task at home and/or at work
Self care (eating, washing, dressing)
Able to perform all self-care activities without problems
Able to perform all self-care activities, albeit with some difficulties
Not able to perform some self-care activities
Not able to perform many self-care activities without help
Not able to perform any self-care activity without help
Leisure activities
Able to perform all types of leisure activities without difficulties
Able to perform all types of leisure activities, albeit with some difficulties
Not able to perform certain types of leisure activities
Not able to perform many types of leisure activities
Not able to perform any type of leisure activities
Pain and/or other complaints
No pain and/or other complaints
Now and then light to moderate pain and/or other complaints
Often light to moderate pain and/or other complaints
Often moderate to severe pain and/or other complaints
Always severe pain and/or other complaints

EXPERIMENTAL DESIGN

The first substantive task the respondents were asked to perform was ranking six health states. Then the respondents were asked to locate the health states on a rating scale, calibrated from 0 (immediate death) to 100 (full health). Full health was defined as the best score on each of the four dimensions. When asked, we told the respondents to imagine that the health states lasted for the rest of their lives. It was explained to the respondents that the intervals between the health states should reflect their strengths of preference: health states that differed slightly in attractiveness should be placed close to each other, whereas health states that differed widely in attractiveness should be placed further apart. Two versions of the questionnaire were used, and the experimental sessions were randomly allocated to one of the two versions. The six health states differed between the two versions of the questionnaire. Health states A, B, C, and D were included in both versions. In addition to these health states, health states E and F were included in version 1 of the questionnaire and health states G and H in version 2. This design allowed us to test whether rating-scale valuations are affected by the other health states included in the assessment task.

Table 2 • Health States A. B. C. D. E. F, G. and H

Health state A
- Able to perform all tasks at home and/or at work without problems
- Able to perform all self-care activities (eating, washing, dressing) without problems
- Able to perform all types of leisure activities, albeit with some difficulties
- Now and then light to moderate pain and/or other complaints

Health state B
- Able to perform all tasks at home and/or at work, albeit with some difficulties
- Able to perform all self-care activities (eating, washing, dressing) without problems
- Unable to participate in certain types of leisure activities
- Often light to moderate pain and/or other complaints

Health state C
- Able to perform all tasks at home and/or at work without problems
- Able to perform all self-care activities without problems
- Able to perform all types of leisure activities, albeit with some difficulties
- No pain and/or other complaints

Health state D
- Unable to perform some tasks at home and/or at work
- Able to perform all self-care activities (eating, washing, dressing), albeit with some difficulties
- Unable to participate in many types of leisure activites
- Often moderate to severe pain and/or other complaints

Health state E
- Able to perform all tasks at home and/or at work, albeit with some difficulties
- Able to perform all self-care activities (eating, washing, dressing) without problems
- Able to perform all types of leisure activities, albeit with some difficulties
- Now and then light to moderate pain and/or other complaints

Health state F
- Unable to perform some tasks at home and/or at work
- Able to perform all self-care activities (eating, washing, dressing) without problems
- Unable to participate in certain types of leisure activities
- Now and then light to moderate pain and/or other complaints

Health state G
- Unable to perform some tasks at home and/or at work
- Able to perform all self-care activities (eating, washing, dressing), albeit with some difficulties
- Unable to participate in certain types of leisure activities
- Often light to moderate pain and/or other complaints

Health state H
- Unable to perform some tasks at home and/or at work
- Able to perform all self-care activities without problems
- Unable to participate in many types of leisure activities
- Often light to moderate pain and/or other complaints

The experiment was designed to test whether context effects would arise both for ‘health states for which the numbers of more-preferred and less-preferred health states varied and for health states for which the numbers of more-preferred and less-preferred health states were constant. One possibility is that the valuation of a specific health state is affected by the number of other health states included in the assessment task that are more or less preferred than the target. It is furthermore possible that the valuation is affected not only by the number of more- or less-preferred health states but also by how much better or worse they are.

To test for consistency of rating-scale valuations, two of the health states varied between group 1 and group 2. In group 1, health states E and F were chosen such that at least one of them (health state E) would be considered more attractive than health state B, both would be considered more attractive than health state D, and both would be considered less attractive than health states A and C. In group 2, health states G and H were chosen such that they would be considered less attractive than health states A, B, and C and more attractive than health state D. Thus, in our experimental design the number of preferred health states varies for health state B and the dependency of the rating scale valuations on the number of preferred health states can be tested by comparing the valuations for health state B between the two groups of respondents. The consistency of the rating-scale valuations for health states for which the numbers of preferred and less-preferred health states are constant can be tested by comparing the rating-scale valuations for health states A, C, and D between the groups. For these health states, the numbers of preferred health states are equal in the two groups, but how much better or worse the additional health states are varies between the groups.

Finally, the respondents were asked to answer standard-gamble questions for health states B and D. The tests reported in this paper were part of a larger experiment, and because the respondents could be asked to perform only a limited amount of tasks in an experimental session, we were forced to confine ourselves to the elicitation of standard-gamble weights for health states B and D. For reasons not related to the present study, the standard-gamble questions differed between the groups. Version 1 respondents answered standard-gamble questions in which the certain options were 10 years in D (Dl0), 30 years in D (D30), and 30 years in B (B30), respectively. Version 2 respondents answered standard-gamble questions in which the certain options were 10 years in B (Bl0), 30 years in B, and 30 years in D, respectively. The outcomes of successful treatment were 10 years in full health in the first question and 30 years in full health in the second and third questions. The outcome of unsuccessful treatment was immediate death in all questions. The respondents were explicitly informed that all profiles would be followed by death. Probability elicitation was by means of a line of values for the probability of successful treatment. Next to this line, a line was drawn with the complementary probability of failure of treatment (immediate death). This display was chosen in an attempt to control for a potential framing bias: only displaying the probability of successful treatment might induce an individual to focus on successful treatment, not sufficiently taking into account the probability of failure of treatment. Psychological evidence of the influence of reference effects on choice is abundant.” The respondents were encouraged to follow a sort of ping-pong strategy. First, they were asked to indieate those values of p, the probability of successful treatment, for which they definitely preferred the certain option; then, those values of p for which they definitely preferred the treatment option (gamble); and, finally, those values of p for which they found it hard to choose between the options. It was pointed out to the respondents, both during the description of the task and during the oral explanation, that they were allowed to indicate a range of values of p for which they found it hard to choose between the options. The midpoint of this range of values was used in the analyses.

STATISTICAL METHODS

Mean values were compared between the two groups by two-tailed two-sample t-tests. The two-sample t-test is robust for non-normality so long as the hypothesis of equal variances in the two samples cannot be rejected. We therefore first tested equality of variances in the two samples by means of an F-test. When equality of variances was rejected, **the** nonparametric Mann-Whitney test was used to analyze the data. The results are presented both for the total sample and for the Stockholm and Rotterdam samples separately. This allows us to test for the stability of the results.

Equation 2 was estimated for the individual data by nonlinear least squares. An alternative to using nonlinear least squares is to take logarithms and to estimate equation 2 by ordinary least squares. However, in that case problems arise when an individual rates a health state to be as good as full health and arbitrary adjustments have to be made. Because this happened in a number of cases, we decided to use nonlinear least squares instead. To examine the appropriateness of equation 2, we included a type of RESET test? to test for functional form. The RESET test amounts to adding the square and higher moments of the predicted values to the model. If the inclusion of these variables leads to a significant improvement in the performance of the model, then there is indication of misspecification. Misspecification indicates that equation 2 does not describe the appropriate relationship between standard-gamble valuations and rating-scale valuations. A disadvantage of the RESET test is that it only indicates misspecification, but it does not indicate which functional form to use instead.

We further tested for heteroskedasticity. The nonlinear least-squares model assumes that the error terms and the independent variable are distributed independently. If this assumption does not hold, the errors are said to be heteroskedastic. Heteroskedas-

Table 3 • Mean Rating-scale Valuations (Standard Errors in Parentheses)*

	Swedish Sample Group 1	Swedish Sample Group 2	Dutch Sample Group 1	Dutch Sample Group 2	Total Sample Group 1	Total Sample Group 2
Health state A	0.8366 (0.0155)	0.6465 (0.0136)	0.8266 (0.0106)	0.8220 (0.0123)	0.8312 (0.0078)	0.8335 (0.0092)
Health state B	0.6420 (0.0190)	0.7404† (0.0187)	0.5922 (0.0185)	0.6916† (0.0157)	0.6151 (0.0134)	0.7145† (0.0123)
Health state C	0.9281 (0.0076)	0.9375 (0.0056)	0.9206 (0.0083)	0.9167 (0.0093)	0.9241 (0.0057)	0.9265 (0.0057)
Health state D	0.4348 (0.0230)	9.4200 (0.0307)	0.3704 (0.0210)	0.4038 (0.0210)	0.4000 (0.0158)	0.4114 (0.0182)
Health state E	0.7530 (0.0148)		0.7400 (0.0136)		0.7460 (0.0100)	
Health state F	0.6208 (0.0243)		0.5500 (0.0192)		0.5825 (0.0156)	
Health state G		0.5808 (0.0270)		0.5418 (0.0197)		0.5599 (0.0165)
Health state H		0.6146 (0.0235)		0.5500 (0.0174)		0.5804 (0.0148)

*See table 2 for descriptions of the health states.
†Significantly different from group 1 at the 1% level.

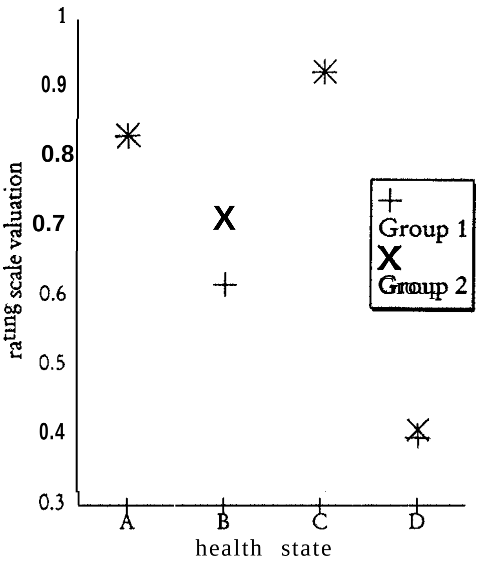

FIGURE 1. **Rating-scale valuations for the two experimental groups (total sample).**

ticity leads to biased estimates of the standard errors of the parameters and thus to the wrong conclusions about the significance of the parameters. We used the Lagrange multiplier tes[24,25] to test for heteroskedasticity.

Results

Table 3 displays the mean rating-scale valuations for the two experimental groups. For health state B, the table shows a marked discrepancy between the two groups. This discrepancy is in the direction hypothesized by the range-frequency model: the valuation for health state B is about 0.10 higher in group 2 compared with group 1. This difference is statistically significant ($p < 0.0011$. The table shows that according to the rating scale, health state B is preferred to two health states in group 1 and to three health states in group 2. Thus, we reject the null hypothesis of no context effect when the number of preferred health states varies. For health states A, C, and D, which are preferred to the same number of health states in both groups, we do not observe significantly different valuations.

Figure 1, summarizes the picture for all four health states for the total sample. Clearly, the rating-scale valuations differ widely only for health state B.

Separate analyses of the Dutch and Swedish samples confirm the above conclusions: for health state B, the difference between the two groups is statistically significant ($p < 0.001$ in both samples); for health states A, C, and D, the difference between the groups is not significant ($p > 0.10$ in all cases).

One might object that the results we obtain are simply an artifact of the difference between the groups. Two remarks are worth making here. First, the respondents were allocated randomly to the two versions. We have no reason to believe that the allocation process introduced any bias. Second, if the groups were different, this should have been reflected in their answers to the other tasks in the experiment. For example, the mean responses to the standard-gamble questions should also have differed between the groups. For comparison, we include the responses to the standard-gamble questions, in table 4. No statistically significant difference was found either for the total sample or for the Dutch and Swedish samples separately.

Even though the above analysis casts doubt on the claim that what is measured by the rating scale is a measurable value function, we decided to estimate equation 2 nevertheless. The results of the estimation procedure are shown in table 5. For the total sample, the estimates of beta differ for B30 and D30, even though their confidence intervals overlap. However, the test statistics indicate severe problems. Both for B30 and for D30, we find indication of heteroskedasticity and misspecification.

Given that the two groups differ systematically in the rating-scale valuations assigned to health state B, it is worthwhile to estimate equation 2 for the two groups separately. Moreover, this allows us to estimate an additional equation, because we can also relate the valuations for D10 and B10, respectively. Table 5 shows that in group 1 the variation in the parameter estimates is less than that in the total sample. Moreover, the test statistic no longer indicates misspecification. We cannot reject the equation proposed by Torrance for this group. However, the Lagrange multiplier test still indicates problems of heteroskedasticity. In group 2, the parameter estimates differ substantially. For B30 and D30 the confidence intervals no longer overlap. Again, there is no indication of misspecification. However, the

Table 4 • Mean Standard-gamble Valuations (Standard Errors in Parentheses)

	Swedish Sample Group 1	Swedish Sample Group 2	Dutch Sample Group 1	Dutch Sample Group 2	Total Sample Group 1	Total Sample Group 2
Health state B	0.8908 (0.0219)	0.8456 (0.0282)	0.8434 (0.0276)	0.8552 (0.0262)	0.8851 (0.0181)	0.8507 (0.0191)
Health state D	0.6923 (0.0326)	0.6318 (0.0395)	0.6667 (0.0388)	0.8846 (0.0403)	0.6784 (0.0282)	0.6597 (0.0256)

Table 5 • Results of the Estimation of Eauation 2

	B30	D30	D10	B10
Total sample				
β	**1.984***	**2.375***		
(Standard error)	(0.103)	(0.140)		
Heteroskedasticity				
$\chi^2(1)$ =	**32.16***	**8.22***		
Misspecification				
$\chi^2(2)$ =	**7.64†**	**11.34***		
Group 1				
β	**2.336***	2.466”	2.683’	—
(Standard error)	(0.169)	(0.207)	(0.226)	—
Heteroskedasticity				
$\chi^2(1)$ =	**7.04***	**7.74***	**5.51†**	—
Misspecification				
$\chi^2(2)$ =	2.76	4.19	2.45	—
Group 2				
β	**1.586***	**2.284***	—	**1.948***
(Standard error)	(0.103)	(0.189)	—	(0.116)
Heteroskedasticity				
$\chi^2(1)$ =	**30.30***	2.01	—	**45.28***
Misspecification				
$\chi^2(2)$ =	1.11	4.55	—	2.32

*p < 0.01; †p < 0.05.

test statistic for heteroskedasticity is significant for **Bl0** and B30.

Discussion

The aim of this study was to test a theoretical foundation of the rating-scale method as it is commonly used in the measurement of health utilities. We have argued that the interpretation generally attached to rating-scale valuations suggests that what is measured is a measurable value function. We have argued that if the rating-scale method indeed elicits a measurable value function, then the rating-scale valuations should be identical in different contexts and should not depend on the other health states included in the rating-scale task. This theoretical property of measurable value functions was tested in an experiment. We found that, contrary to the theory underlying measurable value functions, a rating-scale valuation depends on the numbers of health states that are preferred and less preferred to it. This result is consistent with the findings of Loomes et al.,[19] but contradicts the conclusions of the study by Kaplan and Ernst.‘” Kaplan and Ernst point to the importance of clearly defining the endpoints of a rating scale if context effects are to be avoided (see also Anderson”). It may be that we observed a context effect because the respondents had problems imagining what full health meant. On the other hand, the health states we used consisted of several dimensions, and full health was defined as the health state consisting of the best score on all dimensions. This definition resembles the definition of the best health state in the Kaplan and Ernst study. An alternative reason for the discrepancy between our conclusion and theirs may be a difference in sample sizes. In their study, the experimental groups consisted of **17** respondents, while we had 87 and 85 respondents in the two experimental groups, respectively.-This gives our study a higher probability of detecting true differences.? In fact, for the comparison between the experimental group that valued only attractive health states and the group that valued all health states, Kaplan and Ernst found a pattern similar to that we observed for health state B with approximately the same differences in valuations. That these differences did not reach statistical significance in the Kaplan and Ernst study may have been due to the lower power.

Our study also shows that the severities of the health states had no impact so long as the numbers of preferred and less-preferred health states remained constant. In that case, the rating scale produced reliable valuations. The context effect was manifest only for circumstances in which the numbers of preferred and less-preferred health states varied.

We further presented parameter estimates of the functional form between standard-gamble valuations and rating-scale valuations suggested by Torrance. If the rating scale would indeed elicit a measurable value function, then it should be related to a vNM utility function elicited by the standard gamble. The results of the estimation procedure using individual data were not supportive of a stable relationship: in general, the parameter estimates we obtained differed significantly, we found indications of misspecification, and in most cases there was indication of heteroskedastic errors. Several factors may have contributed to the apparent lack of a stable relationship. The first is that the rating scale may not elicit a measurable value function, which we had already concluded. However, other factors may have entered in as well. For example, the standard gamble may not provide reliable estimates of vNM utilities. Empirical evidence of inconsistencies in expected-utility theory and in standard-gamble valuations is well documented.[22,28] One dominant finding of the literature on violations of expected-utility theory is that individuals do not evaluate probabilities linearly, as expected-utility theory predicts, but weight probabilities. The impact of probability weighting on standard-gamble valuations and the

†For example, given a significance level of 5% and a standard deviation of 0.15 (which was the average standard deviation in our study), our study is able to detect a true difference of 0.10 with 98% probability, whereas Kaplan and Ernst are able to detect this difference with slightly less than 50% probability.

relationship to rating-scale valuations is a topic for future research. Notice, finally, that our approach differed from the approach used by Torrance: Torrance observed the relationship for grouped data, whereas we used the individual data.

It should be emphasized that the results of our study may have been affected by the way we asked the rating-scale and standard-gamble questions and by the health states we used. However, the way we used the methods is common practice in health-utility research, and the health-classification system we selected is frequently used. The rather negative conclusions of this paper, therefore, are generalizable to other studies that have used similar methods and health-classification systems. Whether refinements of the procedures (e.g., spending more time to familiarize respondents with the health states and the tasks) will eliminate the context effects is a topic for future research.

The implications of our study for the use of the rating scale as it is commonly used to measure health utilities are serious. Our results reject the interpretation of rating-scale valuations as points on a measurable value function. Therefore, no theoretical justification for the use of rating-scale valuations in cost-utility analysis can be given from the theory of measurable value functions. This is a serious limitation, because it appears from statements in the literature that the rating scale is intended to measure a measurable value function. Moreover, the sensitivity of the rating-scale valuations to the number of preferred health states casts doubt on the common practice of valuing several health states simultaneously. Studies that have used this method probably have been affected by the context effect shown in this study. Our results suggest that the rating scale may be useful only when health states are valued in isolation. On the other hand, the rating scale has often been advocated because it provides an easy and cheap way to value several health states simultaneously. The main conclusion of our study is therefore negative for the common use of the rating scale in health-utility measurement: the practical appeal of the rating scale comes at the cost of inconsistency in the valuations it elicits.

The authors are grateful to the editor and three anonymous referees for helpful comments on a previous draft, to Maureen Rutten-van Mölken for suggestions with respect to the selection of the health states, and to Jaco van Rijn for assistance in running the experimental sessions.

References

1. Drummond MF, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. Oxford, U.K.: Oxford Medical Publications, 1987.
2. Nord E. The person-trade-off approach to valuing health care programs. Med Decis Making. 1995;15:201-8.
3. Torrance GW, Feeny D. Utilities and quality-adjusted life-years. Int J Technol Assess Health Care. 1989;5:559-75.
4. Broome J. QALYs. J Public Economics. 1993;50:149-57.
5. Richardson J. Cost-utility analysis: what should be measured? Soc Sci Med. 1994;39:7-20.
6. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. Socioeconomic Planning. 1976;10:129-36.
7. Read JL, Quinn RJ, Berwick DM, Finebeg HV, Weinstein MC. Preferences for health outcomes: comparison of assessment methods. Med Decis Making. 1984;4:31.5-29.
8. Hornberger JC, Redelmeier DA, Peterson J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. J Clin Epidemiol. 1992;45:505-12.
9. Torrance GW, Boyle MH, Horwood PH. Application of multiattribute utility theory to measure social preferences for health states. Operations Research. 1982;30:1043-69.
10. Torrance GW. Measurement of health state utilities for economic appraisal: a review. J Health Econ. 1986;5:1-30.
11. Keeney R, Raiffa H. Decisions with Multiple Objectives. New York Wiley, 1976.
12. Krantx DH, Luce RD, Suppes P, Tversky A. Foundations of Measurement. Vol. 1. New York: Academic Press, 1971.
13. Dyer JS, Sarin RK. Measurable multiattribute value functions. Operations Research. 1979;27:810-22.
14. Dyer JS, Sarin RK. Relative risk aversion. Management Science. 1982;28:87.5-86.
15. Arrow KJ. Alternative approaches to the theory of choice in risk-taking situations. Econometrica. 1951;19:404-37.
16. Ellsberg D. Classic and current notions of "measurable utility." Economic J. 1954;64:528-56.
17. Fishburn PC. Retrospective on the utility theory of von Neumann and Morgenstern. J Risk and Uncertainty 1989;2:127-58.
18. Sarin RK. Strength of preference and risky choice. Operations Research. 1982;30:982-97.
19. Loomes GC, Jones-Lee M, Robinson A. What do visual analogue scales actually measure? Mimeo. University of York, York, U.K., 1994.
20. Bakker C, Rutten-van Mölken M, van Doorslaer E, Bennett K, van der Linden SJ. Feasibility of utility assessment by rating scale and standard gamble in ankylosing spondylitis or fibromyalgla. J Rheumatol. 1995;22:1536-43.
21. Bennett K, Torrance GW. Methodologic challenges in the development of utility measures of health related quality of life in rheumatoid arthritis. Controlled Clinical Trials. 1991;12: S118-28.
22. Kahneman D, Tversky A. Prospect theory: an analysis of decision making under risk. Econometrica. 1979;47:263-91.
23. Davidson R, MacKinnon JG. Estimation and Inference in Econometrics. Oxford, U.K.: Oxford University Press, 1993.
24. Breusch TS, Pagan AR. A simple test for heteroskedasticity and random coefficient variation. Econometrica. 1979;47: 1287-94.
25. Godfrey LG. Testing for multiplicative heteroskedasticity. J Econometrics. 1978;8:227-36.
26. Kaplan RM, Ernst JA. Do category rating scales produce biased preference weights for a health index? Med Care. 1983; 11:193-207.
27. Anderson NH. Foundations of Information Integration Theory. New York: Academic Press,. 1991.
28. Hershey JC, Schoemaker PJH. Probability versus certainty equivalence methods in utility measurement: are they equivalent? Management Science. 1985;31:1213-31.