Proportional Heuristics in Time Tradeoff and Conjoint Measurement

PEEP F. M. STALMEIER, PhD, THOM G. G. BEZEMBINDER, PhD,
IVANA J. UNIC, MD

The time-tradeoff (TTO) test is widely used to measure quality of life for different health states. Subjects are asked to equate the value of living a given period in an inferior health state to the value of living a shorter period in good health. Applications of TTOs have been criticized based on the fact that the value of future life duration is taken as the future life duration itself. The authors show that for a health state in which a subject does not want to live longer than a specified amount of time, subjects' responses do not comply with the assumption that the value of the period in inferior health is equated to the value of the shorter period in good health. Actually, preference reversals with respect to such a health state point to the use of a proportional heuristic in the TTO test. Comparisons of the TTO test in these subjects with category scaling and difference measurements also favor a proportional interpretation of the TTO test. In tests based on conjoint measurement, these subjects also appear to use a proportional heuristic. Consequences of the use of the TTO test and conjoint measurement in quality-of-life models are discussed. *Key words*: utility assessment; QALY; conjoint measurement; preference reversals; compatibility effect. **(Med Decis Making 1996;16:36–44)**

In medical decision making, the concept of quality-adjusted life years (QALYs) is of considerable importance. It involves the measurement of utilities of health states and life durations. Let (L,Q) denote living L years in the ("constant") chronic state of health Q, followed by immediate death. The prevailing evaluation U(L,Q) of (L,Q) goes by the multiplicative model

$$U(L,Q) = V(L) \times W(Q)$$

where W(Q) evaluates well-being in state of health Q and V(L) denotes the appreciation of life years. The factor W(Q) constitutes the "quality weight" of V(L). The aim of the TTO test is to assess W(Q).

The TTO test is a popular method to elicit utilities for health states. It was introduced by Torrance et al.[1] and has been applied by several other investigators.[2–4] In the TTO test, the subject or patient is presented with a period Y of inferior health Q, and one elicits the number X of healthy life years, $X < Y$, considered equivalent to (Y,Q).[5] The resulting number X is called the *TTO equivalent*. We consider two possible interpretations of the numbers V(L) that play a crucial role in the assessment of W(Q).

Received May 5, 1994, from the Nijmegen Institute for Cognition and Information (NICI), Department of Mathematical Psychology, University of Nijmegen, Nijmegen, the Netherlands (PFMS, TGGB); and the Institute for Radiotherapy, St. Radboud Academical Hospital, Nijmegen, The Netherlands (IJU). Revision accepted for publication February 8, 1995. Supported by grant NUKC 93-0136 from the Dutch Cancer Society.

Address correspondence and reprint requests to Dr. Stalmeier: Department of Mathematical Psychology, University of Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands.

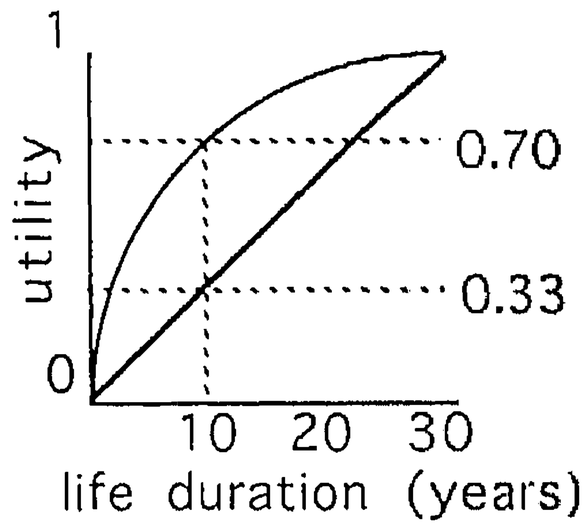

FIGURE 1. Linear and concave appreciations for time duration and their effects on adjustment.

If X healthy life years are equated, i.e., deemed equivalent, to Y years in health state Q, then, according to the multiplicative model, V(X)W(healthy) = V(Y)W(Q). Setting W(healthy) equal to 1, W(Q) equals V(X)/V(Y). In the first interpretation, V(L) is taken as the identity, i.e., V(L) = L. Following this interpretation, W(Q) reduces to X/Y. Thus, the ratio X/Y is taken as a measure W(Q) of the utility of health state Q, and (L,Q) is appreciated by L × W(Q). This interpretation of the TTO test prevails in medical decision making.[1,6]

The second interpretation takes V(L) as a nonlinear function of life years that reflects the common feeling that life years in the near future are more valuable than life years in the distant future. Now W(Q) is taken as V(X)/V(Y). In this case, W(Q) is called an *adjusted* quality weight because it reckons with a nonlinear appreciation of life years. (L,Q) is now appreciated by V(L) × W(Q).[2–4,6,7]

An example shows that the adjustment can be of considerable magnitude. A typical appreciation of life duration is depicted by the curved line in figure 1. Following this curve, the appreciation of living ten

healthy years is 0.7, while that of living 30 healthy years is equal to 1. Now suppose that 30 years in health state Q are considered equal to ten years of healthy life. The quality weight based on V as the identity, shown as the diagonal line in figure 1, is the X/Y ratio 10/30, or 0.33; however, the adjusted quality weight, obtained from the curved line[4] in figure 1, is 0.7, resulting in a positive adjustment of 0.37. Clearly, the discrepancy between the effects of these two interpretations is disturbing because it may lead to different decisions with respect to therapy choice.

In the sequel, we suggest that adjustment is not correct. We report a preference reversal that suggests that a significant proportion of our subjects use a proportional heuristic. Furthermore, we compare the quality weights from the two interpretations of the TTO test with quality weights of other commonly used methods to elicit quality weights for health states, namely category scaling and difference measurement. This comparison also argues for the use of a proportional heuristic. The latter evidence is only circumstantial because there is no "gold standard" with which to compare utilities and because there is no accepted theory to link the quality weights from different elicitation methods. Finally, an ensuing hypothesis concerning the interpretation of conjoint measurement (see below) is tested and confirmed.

PREFERENCE REVERSALS FOR MET HEALTH STATES

We consider the results of the TTO test for a health state in which a subject does not want to live longer than a fixed amount of time, called the "maximum endurable time" (MET).[8] Living with metastasized breast cancer can be such a health state. Women are asked* whether or not they prefer living 25 years with metastasized breast cancer (25,M) to living 50 years with metastasized breast cancer (50,M). Both the (50,M) and (25,M) outcomes are unrealistic because the ten-year survival rate for metastasized breast cancer is only 10%. Here, however, we are interested in the choice behavior, of our subjects. Let us take a closer look at the women who preferred (25,M) to (50,M). In order to be consistent with their preferences for the shorter life duration in the metastasized state, these women should have assigned a smaller TTO equivalent to (50,M) than to (25,M). However, the TTO equivalents turned out to be severely inconsistent with the preference for the shorter life duration and instead complied with a constant proportional tradeoff. We interpret this finding, first reported in Stalmeier et al.[9] as a *preference reversal*, since the TTO equivalent is higher for (50,M) than for (25,M), which suggests a preference for (50,M), contradicting the choice of the preference for (25,M). We interpret these observations as inconsistencies of choice. We argue on the basis of these observed choice inconsistencies that the TTO equivalents do not reflect simple preferences for MET health states. Instead, the subjects use what we will call a *proportional heuristic*. If subjects use a proportional heuristic, then the quality weights should not be adjusted.

COMPARISON WITH CATEGORY SCALING AND DIFFERENCE MEASUREMENT

We calculated two TTO quality weights, adjusted and unadjusted. As explained above, the unadjusted quality weight is the proportion X/Y. The adjusted quality weight is calculated using a function V(L) as measured with 50/50 certainty equivalent gambles for life duration.[4] We compared these two TTO quality weights with quality weights derived from category scaling and from difference measurements. With category scaling, subjects are asked to generate a number between 0 and 10 to express their evaluations of a (50,Q) outcome. This number is a quality weight and is compared with the two TTO quality weights.

With difference measurements,[10] subjects choose the larger of the differences (50,healthy) − (50,Q) and (50,Q) − (1 month,Q). Setting (50,healthy) = 1 and (1 month, Q) = zero, we thus probe whether (50,Q) is appreciated as being closer to (50,healthy) than it is to (1 month, Q), that is, we probe whether W(Q) is valued >0.5 or <0.5. Provided the two TTO quality weights fall on different sides of 0.5, the result of the difference measurement can differentiate between the two TTO quality weights. For instance, if the unadjusted and adjusted TTO quality weights are 0.33 and 0.7 while the difference measurement weight is larger than 0.5, then the latter result forms a plea to use the adjusted TTO weight. If the unadjusted and adjusted TTO quality weights are 0.6 and 0.8 while the difference measurement weight is larger than 0.5, then the difference measurement gives no cue for choosing between the adjusted and unadjusted quality weights.

CONJOINT MEASUREMENT

Conjoint measurement (CM) is a method of representational measurement, and is used here to assess utility. As such, it has recently been applied to problems related to laryngeal cancer and breast cancer.[11–15] In CM, the format of the questions is the same as in the TTO test, for instance, the subject is asked to choose between (Y,Q) and (X,healthy), where $X < Y$. A difference is that in CM (Y,Q) may be compared with (X,Q′), where Q′ is a health state other than "healthy." Another difference is that in CM, the questions for (Y,Q) are interspersed in a random order between questions related to other health states (spaced presentation). With a TTO, for a given (Y,Q), the interviewer adjusts X until the subject expresses indifference; thus, with

*We are indebted to Lia Verhoef for suggesting the selection criterion. It enabled us to detect the preference reversal, albeit that the suggestion was made for another reason.

Table 1 • Numbers of Subjects in the Tests Using the Various Methods

	No. of Subjects	Subjects Preferring (25,M) to (50,M)	Gamble	TTO* Test		Conjoint Measurement		Category Scaling	Difference Measurement
				12 Questions	8 Questions	Long†	Short‡		
Experiment 1	19	16	16	16		8		16	16
Experiment 2	29	17		17			17		
Experiment 3	38	17			38				

*TTO = time tradeoff.
†270 choices.
‡60 choices.

the TTO, the questions concerning (Y,Q) are presented as a block (massed presentation). For a more detailed explanation of CM in medical decision making, see Maas.[14]

Because identical preference questions are asked in TTO tests and CM, it seems plausible to hypothesize that similar heuristics are used in TTOs and CM. A linear relation between the tradeoffs from the two methods would suggest that this is the case.

Method

Three experiments were done to collect the necessary data. Experiment 1 collected extensive data using all methods. In Experiment 2, additional subjects were tested to investigate the preference reversals and the relation between conjoint measurement and TTOs. In Experiment 3, additional data on the preference reversals were obtained in a classroom setting.

EXPERIMENT 1

Subjects

Nineteen women, 20 years old or older, participated in Experiment 1. All subjects were students. Most were majoring in psychology, some in law. Each subject, once selected, received $15.

Procedure

Written health-state descriptions containing the physical, psychological, and social consequences for three health states, namely living with metastasized breast cancer, living after prophylactic mastectomy, and living with genetic counseling, were prepared. The interviews were on an individual basis. In the selection phase, which took about 10 minutes, a subject was told that she participated in a pilot study concerning decision making by women who have an increased risk for breast cancer due to familial history. The subject read the health-state description of living with metastasized breast cancer, and was selected for further participation if she preferred (25,M) to (50,M). In that case, the other two health-state descriptions were handed out. Subjects were asked to read the health-state descriptions carefully at home and imagine as vividly as possible how these health states would affect their personal lives. The test sessions were on three separate days, each session lasting about 50 minutes.

TTO Test

We obtained the number of healthy life years that the subject considered equivalent to Y years in inferior health. This number is the TTO equivalent, denoted by X. The number X is obtained with a bracketing procedure involving forced *choices* between a duration X in perfect health and a fixed longer duration Y in inferior health.[5] After each statement of a choice by the subject, the interviewer changed the value of X, for instance, X was increased if the subject preferred (Y,Q). This was repeated until the subject expressed indifference. The starting number X was chosen randomly within the range of zero to Y years in order to minimize anchoring effects. The subjects were carefully instructed that there were no right or wrong answers and that their answers should reflect their own preferences. The questions were administered on a computer screen.

The TTO test was administered with four durations Y, namely 5, 10, 25, and 50 years, for the three health states, amounting to 12 (L,Q) outcomes. These outcomes were administered in a random order to each subject separately. The TTO test was administered twice over separate sessions. Only the results of the second test are discussed because the first test was meant to familiarize the subjects with the procedure. This last test is denoted by TTO, 12 questions, in table 1.

Gambles for Healthy Outcomes

The subjects were confronted with a choice between living a certain number of years Y and a 50/50 gamble to live either 50 years or 1 month. In all outcomes, the quality of life was healthy. Via a bracketing procedure, Y was varied until the subject expressed indifference. The final number $Y_{0.5}$ is the certainty equivalent. Setting V(50,healthy) = 1 and (1 month,healthy) = 0, we obtain $V(Y_{0.5}) = 0.5$. Next, the certainty equivalent for a 50/50 gamble with $Y_{0.5}$ and one month as outcomes was measured, resulting in the number of years $Y_{0.5}$ with a utility of 0.25. Repeating this procedure with 50/50 gambles, using certainty equivalents obtained earlier as outcomes, certainty

equivalents were measured with utility values of 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, and 0.875. For instance, $Y_{0.625}$ is the certainty equivalent with a utility of 0.625, obtained from the 50/50 gamble with $Y_{0.75}$ and $Y_{0.5}$ as outcomes. A variation of the gamble method (not reported) was also used. The two gamble methods were presented in two separate sessions.

Conjoint Measurement

Conjoint measurement questions were of the form "which do you prefer: 20 years with prophylactic mastectomy or 15 years in genetic counseling?" Life durations are chosen from the ordered set {1 month, 8 years, 16, 20, 23, 26, 29, 32, 42, and 55 years}. Health states were chosen from the ordered set {metastasis, prophylactic mastectomy, genetic counseling, and healthy}. With 10 × 4 attribute levels, 780 different paired comparison are possible. Trivial comparisons such as: "which do you prefer: 23 years in complete health or 8 years with metastasis" can be omitted because the first pair is better on both the health-state and the duration dimensions. Thus, 270 nontrivial paired comparisons remain. These comparisons are presented in a random order[16] on a computer screen. The comparisons are evenly divided over the three sessions. This test is denoted by CM, long in table 1.

Category Scaling

The subjects were asked: "How do you rate the (50,Q) outcome on a scale from 0 to 10?" It was explained that the outcome (50,healthy) had a magnitude of 10 and that the outcome (1 month,healthy) had a magnitude of zero. The category rating is commonly used as a quality weight. Category scaling was done at the end of the third session.

Difference Measurement

In the difference measurement, the subjects were asked: "which difference is the larger: the difference (50,healthy) − (50,Q) or the difference (50,Q) − (1 month,Q)?" The result establishes whether the quality weight for Q is larger or smaller than 0.5. Difference measurements were done at the end of the third session.

EXPERIMENT 2

Subjects

Twenty-nine women, 20 years old or older, participated in experiment. All subjects were students. Most were majoring in psychology. Each subject, once selected, received $5.

Procedure

The interviews were on an individual basis. The test session lasted about 40 minutes. The subjects participated in the TTO test and a shortened CM test, denoted in table 1 by CM short. In the CM test, 60 nontrivial questions were asked with durations and health states from the ordered sets {5, 10, 14, 17, 20, 23, 25 years} and {metastasis, prophylactic mastectomy, healthy}.

EXPERIMENT 3

Subjects

Seventy-six high school students with a mean age of 17.5 years participated. Of these, 38 girls served as subjects, the other half played the role of interviewer.

Procedure

In a classroom setting, two health-state descriptions concerning prophylactic mastectomy and metastasis were read. The TTO questions were written on paper. Four durations, 5, 10, 25, and 50 years, were used, amounting to eight (L,Q) outcomes. The bracketing method was administered by the interviewers. This TTO test is denoted by TTO 8 questions in table 1. After the test, the subjects were asked to indicate whether they preferred (25,M) or (50,M).

Analysis

PREFERENCE REVERSALS FOR MET HEALTH STATES

We determined the number of subjects who preferred the (25,M) to the (50,M) outcome. For these subjects, we determined the numbers of subjects for whom the (50,M) TTO equivalents were larger, equal to, and smaller than the (25,M) TTO equivalents. X/Y ratios for the metastasis outcomes were determined to assess their proportionality.

COMPARISON WITH CATEGORY SCALING AND DIFFERENCE MEASUREMENT

For the TTO data, the unadjusted quality weights were the ratios X/Y. The adjusted quality weights were calculated as follows. For each subject in Experiment 1, the utility function V(L) for the life duration was established by the gamble method. This utility function was fitted with a power, logarithmic, exponential, or logistic function. The best-fitting function, in a least-squares sense, was used as V(L). The adjusted quality weight was equal to V(X)/V(Y).

We compared quality weights for only one (50,Q) outcome. If the adjustment did not lead to a change in the quality weight of more than 0.1, then no result is reported, because a comparison would make little sense given the limited reliability of the methods. Of course, sizeable adjustments were found only for those subjects for whom V(L) was substantially nonlinear. Furthermore, the adjustment was largest for the worst health state, due to a ceiling effect (see figure 1). For these reasons, adjustments are largest for the metas-

Table 2 • Time Tradeoff X/Y Ratios

Group	No. of Subjects	Years with Metastasis 5	10	25	50
NoMET*	21	0.576	0.721	0.715	0.745
Inconsistent	37	0.476	0.557	0.557	0.545
Consistent	13	0.286	0.238	0.138	0.041

*Subjects preferring living 50 years with metastasized breast cancer (50,M) to living 25 years with metastasized breast cancer (25,M). For these subjects, metastasis is not a maximum-endurable-time health state.

tasis health state. Accordingly, that health state is usually chosen for comparison of its two TTO quality weights with the weights from other methods. The difference measurement method determines whether the quality weight of Q is more or less than 0.5. For discriminating between the two TTO quality weights, we note that the two TTO weights can be compared with the difference measurement only if they straddle the 0.5 value point.

CONJOINT MEASUREMENT

In Experiment 1, the (L,Q) outcomes in the CM questions were taken from the sets {1 month, 8 years, 16, 20, 23, 26, 29, 32, 42, and 55 years} and {metastasis, prophylactic mastectomy, genetic counseling, healthy}. A subset of these questions is the same in the TTO test. For the purpose of a comparison between TTO and CM in Experiment 1, it is reasonable to assume that the 25- and 50-year durations in the TTO test have the same utility as the 26- and 55-year durations in the CM test. Consider, for example, the outcome (26,PM). This outcome is paired with each of the outcomes in the set {(23,healthy), (20,healthy), (16,healthy), (8,healthy), and (1 month,healthy)}, named set A. Suppose the subject has a TTO equivalent for (26,PM) of 19 healthy years. In that case, it is natural to assume that in set A, the first two outcomes will be preferred to (26,M) while the last three outcomes will not be preferred. In other words, the set of preferences corresponding with set A will be {1,1,0,0,0}, where 1 (0) means that the corresponding outcome in set A is preferred (not preferred) to (26,PM). In this case, the CM equivalent of (26,PM) was taken as the mean of 16 and 20 healthy years. The CM equivalents were regressed on the corresponding TTO equivalents. We hypothesize that nonlinear associations between these outcomes will not be significant. If the set of preferences corresponding with set A is, e.g., {1,0,1,0,0}, transitivity is violated and no comparison is possible. In the CM test, 55 years was used because the life expectancy of our subjects was about 55 years.

In Experiment 1, eight subjects compared three health states, metastasis, prophylactic mastectomy, and genetic counseling, with (X,healthy). For the 26-year duration as well as the 55-year duration, there were 8(subjects) × 3(health states) = 24 pairs of TTO and CM equivalents. Thus, in Experiment 1, there were 24 × 2(durations) = 48 paired comparisons. In Experiment 2, 17 subjects compared two health states, metastasis and prophylactic mastectomy, each with a duration of 25 years, with (X,healthy). Thus, in Experiment 2, there were 17 × 2 = 34 pairs of TTO and CM equivalents. Therefore, combining Experiments 1 and 2, 82 cases of paired equivalents were analysed. Ten cases were discarded because of transitivity violations.

Results

Table 1 indicates for each experiment how many subjects participated for each measurement method. It follows from this table that a total of 86 subjects participated and that 50 subjects preferred (25,M) to (50,M).

PREFERENCE REVERSALS FOR MET HEALTH STATES

In Experiments 1, 2, and 3, 50 of 86 subjects preferred (25,M) to (50,M). We expected that these subjects would assign lower TTO equivalents to the (50,M) outcome than to the (25,M) outcome. Contrary to our expectation, 37 of the 50 subjects ($p < 0.001$) assigned longer life durations to the (50,M) outcome. We call these 37 subjects *inconsistent*: the data of these subjects entail a preference reversal. We call the other 13 subjects *consistent* because they had lower or equal TTO equivalents for (50,M). Six of these consistent subjects assigned TTO equivalents close to or often equal to zero to all four pairs of life durations with metastasis.

The 36 remaining subjects preferred (50,M) to (25,M). We call these subjects *noMET* because for these subjects, metastasis is not a MET health state.

Of the 36 noMET subjects, 21 (see table 1, Experiment 3, 38 − 17 = 21) participated in the TTO test. These 21 noMET as well as the 37 inconsistent subjects had remarkably stable X/Y ratios for the 10-, 25-, and 50-year metastasis outcomes, as shown in table 2. Disregarding the X/Y ratio for (5,M), the X/Y ratios are constant, as tested by linear and quadratic comparisons. The X/Y ratios for the consistent subjects were significantly lower, as evidenced by a linear comparison ($F(1,12) = 15.45$, $p < 0.002$).

COMPARISON WITH CATEGORY SCALING AND DIFFERENCE MEASUREMENT

For each subject in Experiment 1, we compared the two quality weights from the TTO test, adjusted and unadjusted, with the quality weights from both category scaling and the difference measurement. There were 16 subjects in Experiment 1 for whom all relevant data were available. At the end of Experiment 1, one additional subject changed her mind with respect to

Table 3 • Quality Weights Obtained with Three Methods

	Quality Weight with Time Tradeoff (TTO)		Quality Weight with Other Method*			
	Unadjusted	Adjusted	Category Scaling	Difference Measurement	Outcome‡	Group§
Subject 1	0.26	0.71	0.45 −	>0.5 +	(50,M)	I
Subject 2	0.30	0.69	0.30 −	<0.5 −	(50,M)	I
Subject 3	0.44	0.82	0.40 −	<0.5 −	(50,M)	I
Subject 4	0.59	0.89	0.35 −	<0.5 NA†	(50,M)	I
Subject 5	0.24	0.88	0.40 −	<0.5 −	(50,PM)	C
Subject 6	0.34	0.74	0.50 −	<0.5 −	(50,M)	I
Subject 7	0.16	0.49	0.30 −	<0.5 NA	(50,M)	I
Subject 8	0.54	0.44	0.40 +	<0.5 +	(50,M)	I
Subject 9	0.66	0.76	0.40 −	<0.5 NA	(50,M)	I
Subject 10	0.68	0.86	0.70 −	>0.5 NA	(50,PM)	C
Subject 11	0.48	0.63	0.55 −	>0.5 +	(50,M)	I
Subject 12	0.60	0.78	0.75 +	>0.5 NA	(50,GC)	C
Subject 13	0.84	0.96	0.85 −	>0.5 NA	(50,GC)	X

* −/+ = quality weight from category scaling or difference measurement more/less similar to adjusted TTO than to unadjusted TTO.
†NA = comparison of quality weight from difference measurement with TTO not applicable because TTO weights do not straddle 0.5.
‡M metastasis; PM = prophylactic mastectomy; GC = genetic counseling.
§I = inconsistent; C = consistent; X = noMET (see text).

the selection criterion, which requires preference of (25,M) over (50,M). Her data are also used here. Visual inspection of the utility functions for life duration reveals that 13 are concave, 2 diagonal, 1 convex, and 1 logistic. The latter case is discarded because the TTO quality weights from this subject are either 0 or 1. The two diagonal cases are discarded because adjustment has no effect (see figure 1). One concave case is discarded because the corresponding TTO quality weights are close to 1, also resulting in adjustments that are too small.

Table 3 shows the data for the remaining subjects. For Subject 5, the (50,PM) outcome was chosen because it fulfils all requirements. For Subjects 10 and 12, the TTO equivalents for metastasis are equal or close to zero, leading to too-small adjustments; therefore, other outcomes are chosen. For Subject 13, the metastasis weights are almost equal to the GC weights; therefore, the more realistic (50,GC) outcome is chosen.

The product-moment correlations between quality weights from category scaling and the adjusted and the unadjusted quality weights from TTO are 0.45 (n.s.) and 0.70 ($p < 0.01$), respectively.† The +/− signs in columns 4 and 5 indicate whether the quality weights from the category and difference methods are closer to the adjusted (+) or unadjusted (−) TTO quality weights. The quality weights from the category scaling are closer to the unadjusted TTO values for 11 of 13 subjects ($p < 0.02$).

For the first seven subjects, substantial differences between unadjusted and adjusted quality weights larger than 0.30 were found. For these seven subjects, the corresponding quality weights from the category scaling were all closer to the seven unadjusted TTO quality weights ($p < 0.01$).

As explained in the introduction, the difference method does not differentiate between the two TTO quality weights if these two weights fall on the same side of 0.5. In that case, NA (not applicable) appears in column 7 of table 3. For the first seven subjects in table 3, the difference method also agreed better with the unadjusted TTO quality weights for four of the thus remaining five (n.s.) subjects.

We conclude that the adjusted quality weights from the TTO disagree with the quality weights from the category scaling; the difference measurements tend to disagree as well. The disagreement is pronounced when the adjustment is large, that is, larger than 0.3.

CONJOINT MEASUREMENT AND PROPORTIONAL HEURISTICS

The regression between the conjoint measurement and TTO equivalents shows a correlation of 0.94, ($p < 0.0001$). The regression slope is 0.89, with a standard error of 0.04. As expected, the intercept is not significantly different from zero. A test for deviations of linearity is not significant [$F = 1.819$, $df = (13,57)$, $p < 0.06$]. Note that the linear relation is not an artifact of averaging over subjects, as identical analyses of individual data also show a linear relation between the CM and TTO equivalents. We conclude that our assumption that CM and TTO equivalents are linearly related is not disconfirmed by our data, though the power of the test may not have been sufficient to do so. Nevertheless, the linear association is strong in the sense that it accounts for 88% of the variance..

†Our data show no evidence for a power relation between the category scaling scores and the unadjusted TTO scores.

General Discussion

PREFERENCE REVERSALS FOR MET HEALTH STATES

For a bad health state, metastasized breast cancer, 50 of 86 subjects preferred (25,M) to (50,M). We found in the TTO test that the majority of these subjects, 37 of the 50, wanted more healthy life years for the longer duration of life with metastasized cancer than for the shorter one. Hence, the highest number of healthy years was assigned to the less preferred outcome. This shows that the TTO equivalents do not reflect the preferences of these 37 subjects. These results entail a preference reversal.

It might be the case that subjects initially preferred (25,M) to (50,M), but that this preference changed during the experiment. This could explain the observed preference in the TTO data. However, in Experiments 1 and 2, we carefully elicited the prior preference of (25,M) over (50,M), and all subjects (but one) reaffirmed that preference at the end of the experiment. In Experiment 3, the preferences between (25,M) and (50,M) were elicited *after* the elicitation of the TTO equivalents. Therefore, we conclude that a preference change does not underlie our finding. We confronted seven subjects with their preference reversals. Six of the seven did not want to change their answers at all! The remaining subject did change her answers, but did not change them enough to alleviate the preference reversal. This persistence makes the preference reversal seem genuine.

Our explanation for this preference reversal is as follows.[9] For prophylactic mastectomy and/or genetic counseling, the 50-year outcome is always more attractive than the 25-year outcome. A monotone *heuristic*, that is, a heuristic that equates more healthy life years to the longer duration in inferior health, is plausible for these two health states. If this heuristic is applied to the metastasis health state, this will give rise to the observed choice paradox. The *compatibility principle*[17–20] reinforces this interpretation. It states that "stimulus components that are compatible with the response are weighted more heavily than those that are not." In the TTO task, subjects strongly concentrate on the life-duration dimension. According to the compatibility principle, the life-duration dimension in the TTO question will receive a larger weight, while the weight for the health state will diminish. The joint operation of the monotone heuristic and the compatibility principle is apparently so strong that most subjects do not see that life duration should be evaluated negatively, inasmuch as (25,M) is preferred to (50,M). The monotone heuristic can be specified as a *proportional heuristic*, where, in a TTO question for (Y,Q), a subject chooses X as a proportion of Y. This follows because the X/Y ratios observed for the (25,M) and (50,M) outcomes are the same. It is through the blind use of this proportional heuristic that preference reversals occur. This blind use is driven by the compatibility effect.

Nevertheless, the TTO quality weights X/Y are inconsistent for these subjects in the sense that the X/Y ratios for the (50,M) outcome are the same and not less than the ratios of the (25,M) outcome. Thus, the X/Y ratios do not reflect that these subjects prefer (25,M) over (50,M). Our conclusion is that, for bad health states, one should ask unidimensional preference questions such as "which do you prefer: (5,M) or (10,M)?" When a subject prefers the shorter to the longer durations, then our data show that preference reversals are frequently observed. From the normative viewpoint of a rational preference theory, the TTO test should be abandoned beyond the most preferred duration for those subjects who indeed have a preference reversal. This advise concurs with the discussion in Sutherland et al.,[8] in which they say that "a failure to identify best and worst outcomes for different time frames ... may obscure the existence of a MET ... this could lead to erroneous interpretations of time preference curves. ..."

One may ask whether preference reversals may be found for real patients. We suspect that the answer will be positive, because the proportional heuristic is cognitive: cognitive effects on utility measurement such as framing[21–23] have been shown to be persistent in patients. We are pursuing this matter further with real patients and for more realistic durations.

As an aside, in the TTO task, life duration is varied as a response measure, whereas in the CM task the response is a simple preference. The compatibility effect predicts that subjects will weight the life-duration dimension more heavily in the TTO task as compared with the CM task because in the TTO task, the response is compatible with the life-duration dimension. It is interesting to note that the TTO equivalents were indeed 10% larger than the CM equivalents; this finding also agrees with and confirms the compatibility effect.

COMPARISONS WITH CATEGORY SCALING AND DIFFERENCE MEASUREMENT

For the 16 subjects selected in Experiment 1, the unadjusted quality weights X/Y correlated significantly with the quality weights derived from category scaling. On the contrary, the adjusted quality weights V(X)/V(Y) correlated less strongly with the category scaling weights. The weights with large adjustments in particular, disagreed strongly with the category scaling weights. When confronted with the adjustments, several subjects complained that such large positive adjustments are unrealistic. The difference measurements were also more in line with the unadjusted TTO weights. We conclude that our data strongly suggest that for subjects who prefer (25,M) to (50,M), the X/Y quality weight shows more convergent validity with

category scaling and, to a lesser extent, with difference measurement. This supports our claim that these particular subjects use a proportional heuristic.

HOW WIDESPREAD IS THE PROPORTIONAL HEURISTIC?

The preference reversal in inconsistent subjects and the support for the proportional heuristic in the previous paragraph indicate that the inconsistent subjects use a proportional heuristic. One might argue that our demonstration of deviant results after adjustment in the "Comparison with category scaling and difference measurement" section is flawed because nine of the 13 subjects in Experiment 1 were inconsistent (as registered in the last column of table 3); in that case, the argument continues, in the previous section we selected subjects who use a proportional heuristic to prove our point against adjustment. In this section, we put forward the conjecture that the proportional heuristic is not confined to inconsistent subjects.

The inconsistency of the TTO responses pertains only to our finding that inconsistent subjects prefered (25,M) to (50,M). This does not imply that these subjects used an atypical *heuristic* with the TTO test. Now, for the metastasis health state, the noMET and inconsistent‡ subjects, that is, 73 of 86 subjects, had similarly-shaped TTO curves for the metastasis health state and a *constant proportional tradeoff* for durations longer than 5 years with metastasis (see table 2). Only the 13 consistent subjects' choices showed X/Y ratios that decreased with duration. Therefore, given the similar structures of the TTO responses, as based on the constant proportional tradeoff criterion, we have some reason to believe that the noMET subjects used a heuristic similar to that used by the inconsistent subjects. If this is true, the inconsistent subjects may be used for the "comparisons with category scaling and difference measurement."

ADJUSTMENT, YES OR NO?

Should we adjust the quality weights from the TTO test for the fact that short-term life years are valued differently from long-term life years?[2,4] The use of a proportional heuristic argues against adjustment, provided that the proportional heuristic occurs without a preference reversal. The absence of a preference reversal supports the correctness of the preference theory and thus the QALY models described in the introduction. There, we identified a quality weight of X/Y with no adjustment; V(X)/V(Y) was identified with adjustment. The proportional heuristic leads in a natural way to the use of the ratio X/Y to characterize the quality weight and, thus, forms a plea against adjustment. Therefore, the proportional heuristic without a preference reversal argues against adjustment. Recently, Bleichrodt and Johannesson[24] showed that the ranking of unadjusted TTO quality weights agrees better with rank-ordered health profiles than the ranking of adjusted quality weights. This finding agrees with our plea against adjustment. More data are needed that support the use of a proportional heuristic without a preference reversal.

CONJOINT MEASUREMENT

The proportional heuristic has important consequences for the interpretation of CM. In CM, value functions are derived that model the appreciation of life duration and health states. These value functions depend on the *ordinal* relations in the preferences. Unfortunately, ordinal relations do not uniquely determine value functions: different value functions are strategically equivalent if they preserve the order of preferences over the outcomes.[12] In our case, the value functions from CM are equivalent up to a linear or exponential transform.[15] This raises the problem of pinpointing the correct value function in CM.

A general approach to solving this problem is to investigate the relation between utility scales and value functions derived from preferences. For instance, utility scales derived from gambles were compared with strength-of-preference scales from difference measurements by Barron et al.[25] Maas and Wakker[15] transformed the CM values for duration so as to fit the gamble utility function for life duration. The quality weights for the health states are also affected by this transformation. The method of Maas and Wakker is similar to the adjusted TTO test in the sense that both methods transform preference scales to utility scales via the function V(L), where V(L) denotes the appreciation for life duration and is obtained via the gamble method. In other words, both methods adjust W(Q) via V(X)/V(Y).

We propose a different solution. If one accepts our conclusion that subjects use a proportional heuristic in the TTO test, then the fact that a linear relation explains 88% of the variance between the CM and TTO equivalents suggests that the same heuristic is applied in the CM method. In that case, in terms of the multiplicative model (see the introduction), one should set V(L) = L. Consequently, the quality weights for the health states in CM should be derived with a function V(L) = L for life duration.§

‡The terms inconsistent, consistent, and noMET are introduced in the Results section.

§A strict linear relationship between the CM and TTO equivalents is not possible because: 1) the CM and TTO quality scales coincide at the endpoints death (maximal tradeoff) and perfect health (no tradeoff); and 2) the TTO equivalents are larger than the CM equivalents between the endpoints due to the compatibility effect.

QUALITY-OF-LIFE MODEL

Above, we suggest that a proportional heuristic argues against adjustment. This might be construed to mean that the QALY model for decision making should be L × W(Q), as described in the introduction. However, we want to make a distinction between 1) the *descriptive* model L × W(Q) for interpreting subjects' responses in the TTO test and 2) the *normative* model V(L) × W(Q), to be used for quality-of-life modeling in decision making.

With a proportional heuristic, the ratio X/Y is the natural way to characterize the quality weight. Indeed, this is formally equivalent to using L × W(Q) as a *descriptive* model of the subjects' responses in TTO. Formally, in the descriptive model, using X/Y corresponds with setting V(L) = L in the TTO method. However, this does not mean that life years are never discounted: actually, the appreciation of life duration, as established via the gambles, was nonlinear for 15 of 17 subjects. Our results merely suggest that our subjects did not discount life years in the TTO test.

From a *normative* point of view, discounting should be taken into account[26] in the quality-of-life model as used in a decision tree. We propose the following normative prescription: The duration function V(L) should be elicited with an elicitation method that is sensitive to the discounting of life years such as the gamble method of difference measurement. If the TTO method is chosen to elicit quality weights, then the best characterization of W(Q) is X/Y, as follows from the descriptive model. Finally, in the QALY model for decision making, one appreciates (L,Q) by V(L) × X/Y. Likewise, if CM is used, the quality weights for Q should be derived with a function V(L) = L for life duration.

This normative prescription V(L) × X/Y stands midway between the adjusted V(L) × V(X)/V(Y) and the unadjusted L × X/Y quality of life models. It has the advantage of preventing unrealistic positive adjustments of W(Q) when nearby life years are appreciated more than distant life years because we take the quality weight as X/Y and not as V(X)/V(Y). It also takes into account via V(L) that nonlinear evaluations of life years exist. Once again, the proposed normative prescription presumes that the proportional heuristic is indeed as widespread as we assume.

The authors thank colleagues of the Mathematical Psychology Department and Prof. van Daal of the Radiotherapy Department for helpful discussions. Peter Wakker and Lia Verhoef are especially acknowledged. Frans Gremmen and Ad van der Ven are acknowledged for statistical advise. The authors thank the anonymous referee for detailed criticisms, from which the paper has benefited.

References

1. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res. 1972;7:118–33.
2. McNeil BJ, Weichselbaum R, Pauker SG. Speech and survival: tradeoffs between quality and quantity of life in laryngeal cancer. N Engl J Med. 1981;305:982–7.
3. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. Oper Res. 1980;28:206–24.
4. Miyamoto JM, Eraker SA. Parameter estimates for a QALY utility model. Med Decis Making. 1985;2:191–213.
5. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chronic Dis. 1978;31:697–704.
6. Sox HC, Blatt MA, Higgins MC, Marton KI. Med Decis Making. Boston, MA: Butterworths, 1986.
7. Stiggelbout AM, Kiebert GM, Kievit J, Leer JWH, Stoter G, Haes JCJM de. Utility assessment in cancer patients: adjustment of time trade-off scores for the utility of life years and comparison with standard gamble scores. Med Decis Making. 1994;14:82–90.
8. Sutherland HJ, Llewelyn-Thomas H, Boyd NF, Till JE. Attitudes toward quality of survival. The concept of "maximum endurable time." Med Decis Making. 1982;2:299–309.
9. Stalmeier PFM, Wakker PP, Bezembinder TGG. Preference reversals: violations of unidimensional procedure invariance [manuscript submitted for publication].
10. Dyer JS, Sarin RK. Measurable multiattribute value functions. Oper Res. 1979;20:507–19.
11. Keeney RL, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: Wiley, 1976.
12. von Winterfeldt D, Edwards W. Decision Analysis and Behavioral Research. Cambridge, England: Cambridge University Press, 1986. [See pages 339–340 for a discussion of strategically equivalent scales in conjoint measurement.]
13. Verhoef CG, Maas A, Stalpers LJA, Verbeek ALM, Wobbes Th, van Daal WAJ. The feasabilty of additive conjoint measurement in measuring utilities in breast cancer patients. Health Policy. 1991; 17:39–49.
14. Maas A, Stalpers L. Assessing utilities by means of conjoint measurement: an application in medical decision analysis. Med Decis Making. 1992;12:288–97.
15. Maas A, Wakker P. Additive conjoint measurement for multiattribute utility. J Math Psychol. 1994;38:86–101.
16. Wells AJ. Optimal presentation orders for the method of paired comparisons. Br J Math Stat Psychol. 1991;44:181–93.
17. Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. Psychol Rev. 1988;95:371–84.
18. Fischer GW, Hawkins SA. Strategy compatibility, scale compatibility, and the prominence effect. J Exp Psychol: Hum Percept Perform. 1993;19:580–97.
19. Slovic P, Griffin D, Tversky A. Compatibility effects in judgment and choice. In: Hogarth RM, ed. Insights in Decision Making: A Tribute to Hillel J. Einhorn. Chicago, IL: University of Chicago Press, 1990, pp 5–27.
20. Delquié P. Inconsistent trade-offs between attributes: new evidence in preference assessment biases. Manage Sci. 1993; 39:1382–95.
21. McNeil BJ, Pauker SG, Sox HC, Tversky A. On the elicitation of preferences for alternative therapies. N Engl J Med. 1982; 306:1259–62.
22. Llewellynn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. Med Decis Making. 1982;2:449–62.
23. O'Connor AM. Effects of framing and level of probability on patients' preferences for cancer chemotherapy. J Clin Epidemiol. 1989;42:119–26.
24. Bleichrodt H, Johannesson M. Standard gamble, time trade-off, and rating scale: experimental results on the ranking properties of QALYs [manuscript submitted for publication].
25. Barron FH, von Winterfeldt D, Fischer GW. Empirical and theoretical relationships between value and utility functions. Acta Psychol. 1984;56:233–44.
26. McNeil BJ, Weichselbaum R, Pauker SG. Fallacy of the five-year survival in lung cancer. N Engl J Med. 1978;299:1397–1401.