A Consistency Test of the Time Trade-Off

March 2003

Han Bleichrodt, Erasmus University, Rotterdam, The Netherlands

Jose Luis Pinto, University Pompeu Fabra, Barcelona, Spain.

Jose Maria Abellan-Perpiñan, University of Murcia, Spain.

Address correspondence to: Han Bleichrodt, iMTA/iBMG, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands. Email: bleichrodt@bmg.eur.nl. Internet: www.bmg.eur.nl/personal/bleichrodt

Acknowledgements: Two anonymous referees gave very helpful comments. Han Bleichrodt's research was made possible by a fellowship from the Royal Netherlands Academy of Arts and Sciences and by a grant from the Netherlands Organisation for Scientific Research (NWO). Financial support for the experiments reported in the paper was obtained from the Direccion General de Ciencia y Tecnologia (DGICYT PB 94-0848/95) and from SEC 2000-1087.

Abstract

This paper tests the internal consistency of time trade-off utilities. We find significant violations of consistency in the direction predicted by loss aversion. The violations disappear for higher gauge durations. We show that loss aversion can also explain that for short gauge durations time trade-off utilities exceed standard gamble utilities. Our results suggest that time trade-off measurements that use relatively short gauge durations, like the widely used EuroQol algorithm (Dolan 1997), are affected by loss aversion and lead to utilities that are too high.

KEY WORDS: Cost-Utility Analysis, Time Trade-off, Loss Aversion.

JEL CLASSIFICATION: I10

1. Introduction

This paper studies the consistency of time trade-off utilities. The time trade-off is a widely used technique to elicit health state utilities. The EuroQol algorithm, a frequently employed algorithm to compute health state utilities, is based on time trade-off valuations (Dolan, 1997). Several studies provide empirical evidence that the time trade-off captures individual preferences for health well (van Busschbach, 1994; Dolan et al., 1996; Bleichrodt and Johannesson, 1997). Richardson (1994) and Dolan (2000) give theoretical arguments in favor of the time trade-off.

Inconsistencies in time trade-off measurements were found by Stalmeier in several studies. Stalmeier, Bezembinder, and Unic (1996), Stalmeier, Wakker, and Bezembinder (1997), and Dolan and Stalmeier (2003) observed preference reversals between direct choices and time trade-off judgments for health states of low quality, i.e. health states that are close to or worse than death. They attributed these reversals to a proportional heuristic that people use in answering time trade-off questions. These preference reversals do not occur for health states that are clearly preferred to death.

The common endpoints in time trade-off measurements are full health and death. Stalmeier (2002) found inconsistencies in time trade-off utilities when the endpoints used in the elicitation vary. His findings indicate no problems for time trade-off measurements in which endpoints are held fixed, because he observed that the relative size of utility differences, which is the information used in cost-utility analyses, does not depend on the endpoints used.

The above findings suggest that time trade-off measurements may be problematic for health states close to or worse than death and in analyses in which the endpoints in the

elicitation task vary. Time trade-off measurements that use health states clearly preferred to death and that do not vary the endpoints, which is the common case in cost-utility analysis, appear to be on much firmer ground. The present paper will show, however, that inconsistencies also occur in the latter case. What is worse, these inconsistencies are systematic and cannot be explained by random error. We show that the systematic inconsistency can be explained by loss aversion (Kahneman and Tversky, 1979, Tversky and Kahneman, 1991), the idea that people evaluate outcomes as gains and losses from a reference point and are more sensitive to losses than to equally sized gains. The inconsistencies in the time trade-off decrease with the gauge duration used. This finding has interesting implications for the use of the time trade-off in health utility measurement. It also suggests that the EuroQol algorithm leads to health state utilities that are affected by loss aversion.

Two other recent papers have also performed consistency tests of the time trade-off (Spencer, 2003; Clarke, Wolstenholme, and Johnstone, 2003). Both studies found less evidence of systematic inconsistencies in the direction predicted by loss aversion. These two studies used different designs than ours, however, which may partly explain the difference in findings. We discuss these studies in the concluding section of the paper.

The paper is structured as follows. In Section 2 we describe the consistency test used in the paper. In Section 3 we explain how loss aversion can explain why the time trade-off might violate the consistency test in a systematic manner. Section 4 describes the design and results of two experiments that test the consistency of time trade-off measurements. Section 5 concludes the paper.

2. The consistency test

Let (T,X) denote T years in health state X. The conventional procedure to elicit the time trade-off utility of a health state A is to specify some gauge duration T_1 in A and to ask a client, a patient or a member of the general population, to specify the number of years T_2 in full health (FH) so that he is indifferent between (T_1,A) and (T_2, FH). The time trade-off utility of A is then computed as $\frac{T_2}{T_1}$.

Even though the conventional procedure is standard in time trade-off measurements, we might as well measure the utility of health state A through an alternative procedure in which the number of years in full health is specified and a client is asked for the equivalent number of years in A. The elicited time trade-off utility should be independent of the procedure used. Otherwise, we would end up with two different time trade-off utilities for health state A and no grounds to prefer one utility over the other.

The consistency test we performed was based on the above argument and amounted to the following. In a first round of experimental questions, described in Section 4, we asked participants to answer the conventional time trade-off question, i.e. to state the number of years in full health that they considered equivalent to T_1 years in health state A. Suppose a participant indicates that he is indifferent between T_1 years in A and T_2 years in full health. Then we asked him in a second round of experimental questions to state the number of years T_1' in A that he considers equivalent to T_2 years in full health, where T_2 was substituted from the first round. The time trade-off method implies that $T_1' = T_1$ except for random error.

It is well known that the time trade-off assumes linear utility for duration. This is a restrictive assumption and several authors have proposed to adjust time trade-off utilities for time preference (Johannesson, Pliskin, and Weinstein, 1994, Dolan and Jones-Lee, 1997). It is important to note that our consistency test does not depend on the assumption of linear utility. All that is required is that indifference is symmetric (i.e. for all a,b, a ~ b if and only if b ~ a) and that participants prefer more years to less both in A and in full health. If these requirements are satisfied then the equality $T_1' = T_1$ should hold regardless of the shape of the utility function for duration.

Several studies have shown that there exist health states of low quality for which there exists a maximal endurable time: a duration beyond which additional life-years are valued negatively (Sutherland et al., 1982, Dolan and Stalmeier, 2003, Spencer, 2003). For such health states more years are less preferred and, therefore, our consistency test is not valid.

3. Loss aversion

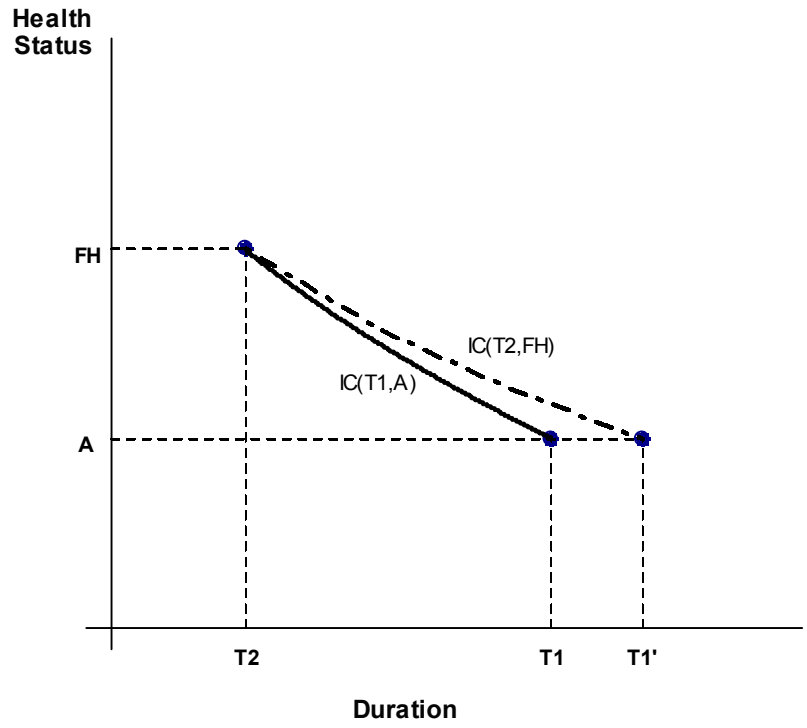
Bleichrodt (2002) argued that two other factors, besides, utility curvature, affect time trade-off utilities: scale compatibility and loss aversion (see also Spencer, 2003). Scale compatibility asserts that respondents tend to give more weight to attributes that are consistent with the response scale used in the elicitation (Tversky, Sattath, and Slovic, 1988). In the time trade-off, the response scale is duration and scale compatibility implies that respondents will focus more on duration than on health status when replying to time trade-off questions. In our consistency test, the same response scale, duration, is used in both stages and scale compatibility, therefore, cannot explain failures of the consistency

test. Loss aversion, however, allows for failures of the consistency test as we explain below.

A formal theory of loss aversion was presented in Tversky and Kahneman (1991). Tversky and Kahneman argue that a person's preferences depend on his reference point. Variations in the reference point will generally lead to different preferences. Tversky and Kahneman additionally assume that people are *loss averse*, i.e. that losses have more impact on preferences than similar sized gains.

In the first stage of the consistency test, a client is given the outcome (T_1,A) and is asked to specify the number T_2 of years in full health that makes him indifferent between (T_1,A) and (T_2,FH). A loss averse client will take (T_1,A) as his reference point and will determine T_2 so that the loss in duration from T_1 to T_2 is exactly offset by the gain in health status from A to full health. Figure 1 illustrates this indifference. $IC_{(T1,A)}$ denotes the indifference curve on which (T_1,A) and (T_2,FH) lie when (T_1,A) is the reference point.

Figure 1: The Impact of Loss Aversion on the Time Trade-off Utilities



In the second stage, the client is given (T_2,FH) and is asked to state the number T_1' of years in A that he considers equivalent to (T_2,FH). That is, the client's reference point shifts to (T_2,FH) and he will now determine T_1' so that the loss in health status from full health to A is exactly offset by the gain in duration from T_2 to T_1'. Because the client is loss averse, losses loom larger than gains and judged from (T_2,FH), (T_2,FH) will now be strictly preferred to (T_1,A). This happens because the difference between T_1 and T_2, which was a loss in the first stage, now becomes a gain and thus by loss aversion gets less weight and the difference between A and full health, which was a gain in the first stage

now becomes a loss and hence gets more weight. Thus, by comparison with (T_2,FH), the positive side of (T_1,A), the difference between T_1 and T_2, gets less weight and the negative side, the difference between A and full health, gets more weight following the shift of the reference point. Hence, by comparison with (T_2,FH), (T_1,A) appears less attractive than in the first stage and, because (T_1,A) and (T_2,FH) were indifferent in the first stage, (T_2,FH) will be strictly preferred in the second stage. Figure 1 illustrates the above argument. The shift in the reference point from (T_1,A) to (T_2,FH) makes the indifference curves more shallow. The new indifference curve is shown as $IC_{(T2,FH)}$. To restore indifference, T_1' must exceed T_1 and thus loss aversion predicts that the second-stage time trade-off utility $\frac{T_2}{T_1'}$ will be lower than the first-stage time trade-off utility $\frac{T_2}{T_1}$.

4. Experiments.

First Experiment

Design

The participants were fifty-one economics students at the University Pompeu Fabra, Barcelona. They were paid five thousand Pesetas (approximately 30 Euro). The experiment was carried out in two personal interview sessions. The two sessions were separated by two weeks. Prior to the actual experiment, we tested the questionnaire in several pilot sessions.

The health state we selected was back pain. We chose back pain because it is a fairly common health problem and participants were likely to know people suffering from it. To describe back pain we used the Maastricht Utility Measurement Questionnaire (Rutten-van Mölken et al., 1995), a slightly adjusted version of the McMaster Health

Utility Index. Table 1 shows the description of back pain. Full health was defined as no limitations on any of the four dimensions.

Table 1: The Description of Back Pain

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

In the first experimental session, the first stage of the consistency test was carried out. We asked five conventional time trade-off questions with the gauge duration of back pain fixed at 13, 19, 24, 31, and 38 years, respectively. We deliberately selected stimulus values that were no multiples of five. The pilot sessions showed that people have a tendency to respond in round numbers, e.g. multiples of five. Our selection of gauge durations intended to make this heuristic less salient. We also learnt from the pilot sessions that participants found it hard to perceive living for very long durations which exceed their life-expectancy. Therefore, we used durations that were substantially lower than participants' life-expectancy. To avoid order effects, we varied the order in which the time trade-off questions were asked.

Recruitment of participants took place one week before the actual experiment started. At recruitment, participants were handed a practice question. Participants were asked to answer this practice question at home. This procedure intended to familiarize participants with the time trade-off questions. Before we started the actual experiment, participants were asked whether they had experienced any problems in answering the

practice question. Participants were then asked to explain their answer to the practice question. This procedure allowed us to test whether participants understood the time trade-off task. In case we were not convinced that a participant understood the task, we explained it again until we were convinced that he understood the task.

Appendix 1 shows the formulation of the time trade-off questions. Indifferences were elicited by a sequence of choices, starting with extreme durations and converging to the duration for which the subject was indifferent. Bostic, Herrnstein, and Luce (1990) showed that a such a choice-based elicitation procedure is less likely to lead to inconsistencies in people's preferences than a matching procedure. Participants reported their answers by filling in a table. At any time during the interview, participants were allowed to check earlier responses and to adjust these if desired. To try and avoid response errors, the participants were asked to confirm the elicited indifference value after each question. The final comparison was displayed once again and participants were asked whether they agreed that the two options were equivalent. In case they did not agree, we would repeat the elicitation procedure for that question.

In the second session, the second stage of the consistency test was carried out. The indifference values for each of the five first-stage questions were substituted and the equivalent number of years with back pain was elicited. Experimental procedures were similar to the first stage. The experiment was part of a larger experiment. The presence of the other experimental tasks and the delay of two weeks between the experimental sessions make it unlikely that participants would recall their previous answers and would note the relationship between the two sessions.

Differences between first-stage and second-stage time trade-off utilities were tested both by the paired t-test and by the nonparametric Wilcoxon signed ranks test. We only report the results of the two tests separately if they yield different conclusions.

Our experiment also permits two tests of constant proportional trade-offs, a central assumption underlying time trade-off measurements. Constant proportional trade-offs implies that the time trade-off utility is independent of the gauge duration in the elicitation. Several studies find support for constant proportional trade-offs (Pliskin, Shepard, and Weinstein, 1980, Hall et al., 1992, Bleichrodt and Johannesson, 1997, Stalmeier, Wakker, and Bezembinder, 1997, Dolan and Stalmeier, 2003). Sackett and Torrance (1978) found negative evidence.

Some authors have suggested that support for constant proportional trade-offs is at least partly due to a proportional heuristic (e.g. Dolan and Stalmeier, 2003). The proportional heuristic asserts that in conventional time trade-off question (A,T_1) versus (FH,T_2), people tend to choose T_2 as a proportion of T_1, because it facilitates their choices. If people adopt this heuristic consistently then constant proportional trade-offs will be satisfied. In contrast with most earlier studies of constant proportional trade-offs, we did not use multiples of ten as the gauge duration. Therefore, if support for constant proportional trade-offs is indeed due to a proportional heuristic then we should expect to find less support for constant proportional trade-offs in our study where the proportional heuristic is less easy to apply.

Constant proportional trade-offs could be tested by comparing the conventional time trade-off utilities with each other and by comparing the alternative time trade-off

utilities with each other. Significance of differences was tested both by analysis of variance and by the nonparametric Friedman test.

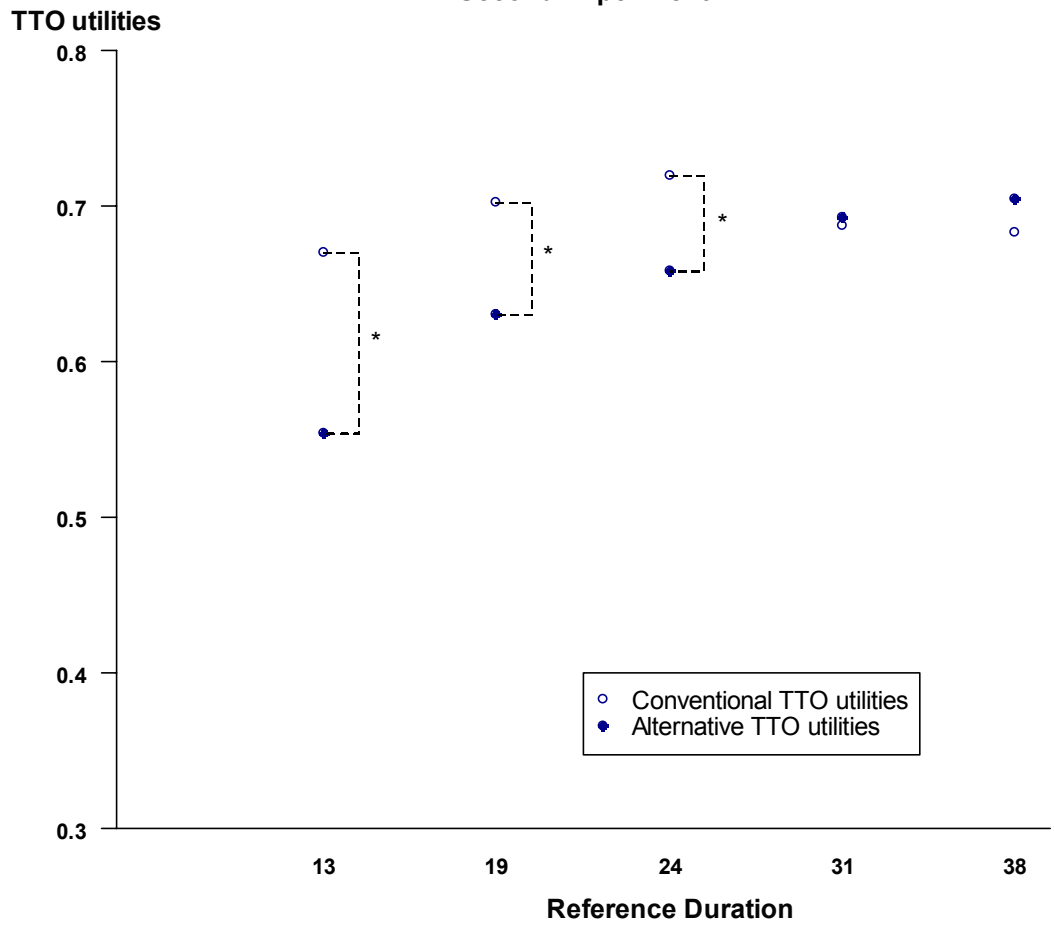
Results

We excluded two subjects from the analyses because they were unwilling to make some tradeoffs. Seven other subjects were excluded because their choices implied that they did not always prefer more life-years to less. Hence, the results reported below are based on the responses of forty-two subjects.

Figure 2 displays the conventional and the alternative time trade-off (TTO) utilities. The figure shows that the time trade-off fails the first three consistency tests (those in which the gauge duration in the first stage of the experiment is equal to 13, 19, and, 24 years respectively). The failure is in the direction predicted by loss aversion. In the other two tests (in which the gauge duration in the first stage of the experiment is equal to 31 and 38 years respectively) we found no significant difference between the conventional and the alternative time trade-off utilities.

The findings on constant proportional trade-offs are generally negative. In the first test, the comparison between the conventional time trade-off utilities, we can reject constant proportional trade-offs by the Friedman test ($P < 0.01$), but not by analysis of variance ($P > 0.10$). In the second test, constant proportional trade-offs is rejected both by analysis of variance and by the Friedman test ($P < 0.01$).

Figure 2: Comparison Between Conventional and Alternative TTO Utilities Second Experiment



* denotes significantly different at alpha = 0.01

Second experiment

Background

The second experiment served two purposes. First, it aimed to test the robustness of the findings from the first experiment in a new subject population and using different health states. Second, it aimed to compare conventional and alternative time trade-off utilities with standard gamble utilities. The results from the first experiment suggest that

time trade-off utilities are affected by loss aversion, in particular for short gauge durations. Bleichrodt (2002) argued that loss aversion leads to an upward bias in conventional time trade-off utilities. The question is how serious this bias is. The upward bias due to loss aversion may be useful to correct other biases in time trade-off utilities (Bleichrodt 2002). We included the standard gamble questions to get some insight into the extent to which the bias due to loss aversion is problematic. Many studies indicate that standard gamble utilities are biased upwards (Hershey and Schoemaker, 1985, Bleichrodt, Pinto, and Wakker, 2001). If the upward bias due to loss aversion would lead to conventional time trade-off utilities that exceed standard gamble utilities then this indicates that the bias is problematic**.**

The comparison between conventional and alternative time trade-off utilities and standard gamble utilities was partly motivated by the finding of Dolan et al. (1996) that time trade-off utilities exceeded standard gamble utilities contrary to the common observation that standard gamble utilities exceed time trade-off utilities. Dolan et al could not give a convincing explanation for this contrast in findings (Dolan, 2001). The results from our first experiment, however, suggest that loss aversion can explain these findings. Dolan et al used a relatively short gauge duration of ten years in their time trade-off measurements. Hence, the results from the first experiment suggest a relatively strong upward bias by loss aversion. If loss aversion indeed explains why Dolan et al. found that time trade-off utilities exceed standard gamble utilities, then we should expect to replicate their finding for short gauge durations, where the data from the first experiment suggested an important effect of loss aversion, but not for longer gauge durations, where the first experiment suggested that loss aversion was less important.

Design

Participants were sixty-five economics students from the University of Murcia. They were paid six thousand Pesetas (approximately 36 Euro). The experiment was carried out in small group sessions with at most six subjects per group. Each participant attended three experimental sessions, one for the standard gamble questions, one for the conventional time trade-off questions, and one for the alternative time trade-off questions. The sessions were separated by at least one week. The time gap between the time trade-off questions was two weeks, as in the first experiment. Prior to the actual experiment, the questionnaire was tested in several pilot sessions using university staff as participants.

We used the EQ-5D health states 22122 and 22322. These states are described in Table 2. Throughout the experiment, the health states were labeled health state A and health state B.

Table 2: The Descriptions of Health States A and B

Health State A
Some problems walking about
Some problems with performing self care activities (e.g. eating, washing or dressing)
No problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
Moderate pain or discomfort
Moderately anxious or depressed

Health State B

Some problems walking about
Some problems with performing self care activities (e.g. eating, washing or dressing)
Unable to perform usual activities (e.g. work, study, housework, family or leisure activities)
Moderate pain or discomfort
Moderately anxious or depressed

Experimental procedures were similar to those of the first experiment: the order of the questions was varied, for each experimental task participants received a question to take home, they had to explain their answer to this question before the actual experiment started, and indifferences were elicited through a sequence of choices. A difference with the first experiment was that in each session they got one additional practice question.

In the first experimental session the standard gamble questions were administered. Participants answered six standard gamble questions, three for each health state. Participants were faced with choices between Y years in health state A for certain versus a risky treatment giving probability p of Y years in full health and probability 1−p of immediate death. The starting value of p was set equal to 0.5 in each question. The selected gauge durations for Y were 13, 24, and 38 years.

In the second experimental session, participants answered the conventional time trade-off questions, in the third session the alternative time trade-off questions. As in the standard gamble questions, the selected gauge durations were 13, 24, and 38 years. The experiment was again part of a larger experiment. Hence, recall bias is unlikely to have affected the results.

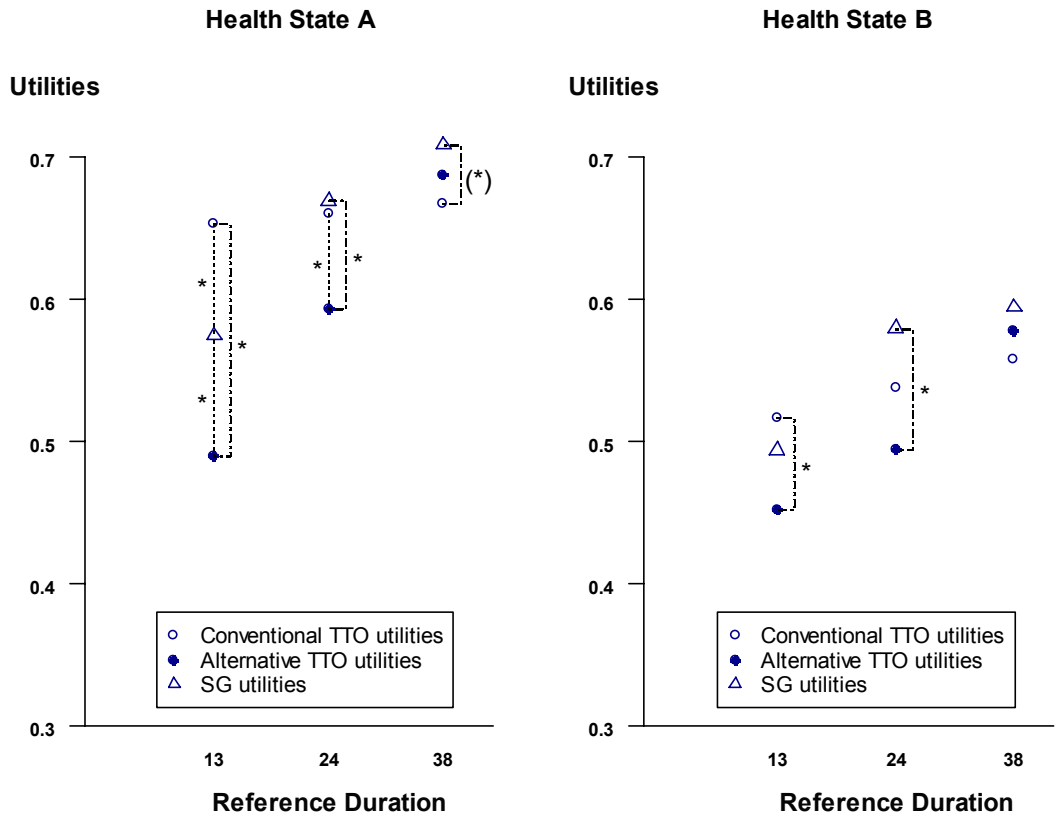
The second experiment yielded not only four tests of constant proportional trade-offs (two for each health state) but also two tests of utility independence of quality of life from life duration (denoted utility independence for short henceforth). Utility independence is a central assumption underlying standard gamble measurements. It implies that the utility score from a standard gamble does not depend on the value at which duration is held fixed. If utility independence does not hold then the standard gamble cannot be used to determine the utility of a health state independent of duration. Bleichrodt and Johannesson (1997) obtained somewhat negative findings on utility independence.

Utility independence could be tested by comparing the three standard gamble utilities that were elicited for health state A with each other and by comparing the three standard gamble utilities that were elicited for health state B with each other. Significance of differences was tested both by analysis of variance and by the Friedman test.

Results

We excluded two and nineteen subjects from the analysis of health state A and B, respectively, because their choices implied that they did not always prefer more life-years to less. This left sixty-three and forty-six participants in the analysis of health states A and B respectively. More subjects had to be excluded for health state B, because B is a worse state than A. The worse a health state, the more likely it is that there is a duration beyond which subjects do not prefer additional life-years.

Figure 3: Comparison Between Conventional TTO, Alternative TTO, and SG Second Experiment



* denotes significantly different at alpha = 0.01 both by the paired t-test and by Wilcoxon's test
(*) denotes significantly different at alpha = 0.01 by Wilcoxon's test only

Figure 3 displays the results. For both health states, the time trade-off fails the consistency test when the gauge duration equals 13 years. For a gauge duration of 24 years, the time trade-off fails the consistency test at a significance level of 1% for health state A and at a significance level of 5% for health state B. In all these four tests, the discrepancy between first-stage and second-stage time trade-off utilities is in the direction predicted by loss aversion. For both health states, the time trade-off satisfies the consistency test, in the sense that we observe no significant difference between

conventional and alternative time trade-off utilities, when the gauge duration is equal to 38 years.

For a gauge duration of 13 years, we are able to replicate Dolan et al.'s (1996) finding that conventional time trade-off utilities exceed standard gamble utilities. For longer gauge durations we find the more common pattern that standard gamble utilities exceed time trade-off utilities. These findings show that Dolan et al.'s findings are consistent with an upward bias due to loss aversion in the conventional time trade-off utilities. The standard gamble utilities always exceed the alternative time trade-off utilities.

For health state A, the difference between conventional time trade-off utility and standard gamble utility is significant at the 1% level for a gauge duration of 13 years. For a gauge duration of 38 years, the difference is significant at the 1% level by the Wilcoxon test (P=0.006) but only at the 5% level by the paired t-test (P=0.022). We indicated this divergence in the figure by putting an asterisk in parentheses. The difference between alternative time trade-off utility and standard gamble utility is significant at the 1% level for gauge durations 13 years and 24 years,.

For health state B, the difference between conventional time trade-off utility and standard gamble utility is significant at the 5% level for a gauge duration of 24 years by the paired t-test (P=0.023) but not by the Wilcoxon test (P=0.053). For a gauge duration of 38 years, the difference is significant at the 5% level by the Wilcoxon test (P=0.050) but not by the paired t-test (P=0.062). The difference between alternative time trade-off utility and standard gamble utility is significant at the 1% level for a gauge duration of 24

years. For a gauge duration of 13 years, the difference is significant at the 5% level by the Wilcoxon test (P=0.017) but not by the paired t-test (P=0.083).

The findings on constant proportional trade-offs are mixed. For health state A, the comparison of the conventional time trade-off utilities supports constant proportional trade-offs ($P > 0.05$). However, constant proportional trade-offs is rejected for the comparison between the alternative time trade-off utilities ($P < 0.001$). For health state B, we cannot reject constant proportional trade-offs for the comparison of the conventional time trade-off utilities by the Friedman test ($P = 0.085$), but we can by analysis of variance ($P = 0.034$). For the alternative time trade-off utilities, constant proportional trade-offs is rejected both by analysis of variance and by the Friedman test ($P < 0.001$). Utility independence is clearly rejected for both health states ($P < 0.001$ in both cases).

5. Conclusion

Main findings

We find inconsistencies in time trade-off utilities. These inconsistencies are in the direction predicted by loss aversion, but arise only when the gauge duration in the time trade-off is relatively low. For longer gauge durations, the time trade-off utilities are consistent, in the sense that they do not depend on the elicitation method. These findings appear robust with respect to the health state used. We are able to replicate Dolan et al's (1996) finding that for short gauge durations conventional time trade-off utilities exceed standard gamble utilities. For longer gauge durations standard gamble utilities exceed conventional time trade-off utilities as is commonly observed in the literature. These

results are consistent with the hypothesis that loss aversion was the cause of Dolan et al.'s finding.

Our findings on constant proportional trade-offs are rather negative. We find mixed evidence on constant proportional trade-offs in conventional time trade-off measurements. In alternative time trade-off measurements, constant proportional trade-offs is violated. Utility independence is violated in both tests that we performed.

Explanations

An explanation why the difference between conventional and alternative time trade-off utilities decreases with duration can be that duration and health status become closer substitutes for higher durations. Several studies have shown that the effect of loss aversion decreases when attributes become closer substitutes (Ortona and Scacciati, 1992, Chapman, 1998). McNeil, Weichselbaum, and Pauker (1981) found that that health status and duration became closer substitutes for higher durations. They observed that people are unwilling to trade life duration for health status if duration is low. That is, for low durations preferences are lexicographic. If duration increases beyond a certain duration, people are willing to give up life duration for improved health status and this willingness increases with duration (see also Pliskin, Shepard, and Weinstein, 1980, Miyamoto and Eraker, 1988).

Our findings on constant proportional trade-offs are to some extent consistent with the proportional heuristic. As expected under the proportional heuristic, we find less support for constant proportional tradeoffs than in other studies that used multiples of ten as gauge durations. Nevertheless, we find some support for constant proportional trade-

offs in the conventional time trade-off measurements. The clear violations of constant proportional trade-offs in the alternative time trade-off measurements suggest that the proportional heuristic plays no role there.

Possible objections

An objection against our study is that in both experiments we elicited the conventional time trade-off before the alternative time trade-off. This may have led to an order effect if people had no clearly defined preferences before coming to the experiment, but constructed their preferences during the elicitation task and benefited in the second session from their experience in the first session. We took some care to avoid the problem of preference construction. In both experiments, subjects received practice questions at recruitment. These questions were intended to induce subjects to think about trading-off life-years against health status.

We are inclined to believe that our results are not seriously affected by an order effect. If the problem of preference construction occurred then it is less likely to have affected the results of the second experiment, because in the second experiment subjects answered the standard gamble questions first. They, therefore, already had opportunity to construct their preferences regarding the trade-off between life-years and health status before they answered the conventional time trade-off questions. The results from the second experiment were, however, similar to those from the first experiment. Moreover, it is hard to conceive of a systematic bias arising from an order effect. If anything, we would expect preferences to be less precise in the conventional time trade-offs, but not systematically biased. We observe, however a systematic difference between

conventional and alternative time trade-off utilities. That having said, it would clearly have been better to include both conventional and alternative time trade-off questions in the first experimental session.

Another possible objection against our study is that we used a young population of students and that it is not clear whether our results can be generalized to the population at large. It is plausible that older people value remaining life duration differently from younger people. Such criticism emphasizes the need to try and replicate our findings in a more representative group of participants. While we agree with the need to replicate our findings, we do not consider the unrepresentativeness of our sample to be an important problem. Many studies show that health state valuations are robust and do not depend in a significant way on the representativeness of the study sample (see de Wit, van Busschbach, and de Charro, 2000 for a review). An indication that our results are robust is that, in spite of the unrepresentativeness of our sample, we were able to replicate the finding by Dolan et al., who used a representative sample, that for short gauge durations conventional time trade-off utilities exceed standard gamble utilities. For longer gauge durations we find the common pattern that standard gamble utilities exceed conventional time trade-off utilities.

Other studies

As noted in the introduction, two other recent studies also examined the consistency of time trade-off measurements. Spencer (2003) found mixed evidence: in one test the time trade-off measurements were consistent, in the other test there were inconsistencies in the direction of loss aversion. Clarke, Wolstenholme, and Johnstone

(2003) found no evidence of systematic inconsistencies in the direction predicted by loss aversion.

Spencer (2003), like us, used a series of choices to elicit indifference durations. The conventional and the alternative time trade-off measurements in her study were not linked, however, and, as Spencer explains, besides loss aversion, time preference, scale compatibility, and maximal endurable time affect the difference between conventional and alternative time trade-off utilities. Moreover, these factors exert opposing influences on the difference between conventional and alternative time trade-off utilities. In our test, time preference and scale compatibility do not affect the results, as we have explained before, and we corrected for maximal endurable time by deleting those subjects who did not satisfy monotonicity with respect to life-years.

Clarke, Wolstenholme, and Johnstone (2003) used a discrete choice experiment in which each subject got only one choice. The most plausible reason for the difference in findings between our study and that of Clarke et al. is the difference in elicitation method. Even though we used a series of choices to elicit indifference, our procedure is closer to matching than Clarke et al.'s who used just one choice. It is well known that people use different evaluation processes in choice tasks than in matching tasks (Tversky, Sattath, and Slovic, 1988). Perhaps, loss aversion is more important in matching than in choice.

Implications

Many practical studies use relatively short gauge durations and our results suggest that the resulting time trade-off utilities are affected by loss aversion. For example, the widely used EuroQol algorithm is based on time trade-off questions that used a gauge

duration of ten years (Dolan, 1997). Based on our findings we therefore have reason to believe that the EuroQol algorithm is affected by loss aversion. This belief is sustained by the fact that we were able to replicate Dolan et al. (1996)'s findings for a short gauge duration, but not for longer gauge durations.

The question is then whether we should strive to avoid the effect of loss aversion on time trade-off utilities? We are inclined to answer this question in the affirmative and to consider loss aversion a bias that should be avoided in health utility measurement. Health utility measurement yields inputs for economic evaluation and medical decision making. The aim of economic evaluations and medical decision making is to help policy makers and patients to make better decisions. That is, economic evaluation and medical decision making are prescriptive techniques and health utility measurement serves to yield inputs for prescriptive decision making. A crucial requirement for prescriptive decision making is that the results of the decision process should not depend on the method that was used to generate the utilities. Equivalent ways to elicit health state utilities should give the same results. Our consistency tests examined this requirement for time trade-off utilities. As noted, we found that the time trade-off only satisfies the requirement for longer gauge durations.

On the other hand, loss aversion is probably not the only bias that affects the time trade-off. An indication that other factors are also at work is that even though the effect of loss aversion (and hence the upward bias in conventional time trade-off measurements) is strongest for short reference durations, conventional time trade-off utilities are not significantly higher for shorter reference durations. Bleichrodt (2002) argued that some effect of loss aversion on the time trade-off utilities may be desirable to offset other

biases in time trade-off measurements. The question of how much loss aversion to allow is not easy to answer. Future research should aim to identify and quantify the biases in time trade-off measurements. We hope that the results from this paper will be helpful in designing such future work.

We should simultaneously strive for the development of new utility measurement instruments. Ideally, economic evaluation and medical decision making should use utility elicitation techniques that are not susceptible to biases such as loss aversion.

Finally, several studies have suggested that there exists a concave relationship between the standard gamble and the time trade-off and that it might be possible to obtain standard gamble utilities by adjusting time trade-off utilities for risk attitude (Miyamoto and Eraker, 1985, Stiggelbout et al., 1994). The results from our paper suggest that the relationship between time trade-off utilities and standard gamble utilities is complex and depends, among other things, on the gauge duration used.

Appendix 1: Formulation of the Back Pain Questions

Suppose that you have 13 more years to live with back pain. In this question you are asked to state the number of years in full health that you consider equivalent to living for 13 more years with back pain. That is, you have to determine the number Y that makes the following two options equivalent:

1. Living for 13 years with back pain. After these 13 years you die.
2. Living for Y years in full health. After these Y years you die.

Use the following table to answer this question.

	Your current situation is 1	You can change to situation 2	DECISION		
Step	Years with back pain	Years in full health	I remain in 1	I am indifferent between 1 and 2	I change to 2
1	13	13			
2	13	0			
3	13	11			
4	13	2			
5	13	9			
6	13	4			
7	13	7			
8	13	5			

References:

Bleichrodt, 2002. A New Explanation for the Difference Between Standard Gamble and Time Trade-Off Utilities. Health Economics 11, 447-456.

Bleichrodt, H. and Johannesson M., 1997. Standard Gamble, Time Trade-Off, and Rating Scale: Experimental Results on the Ranking Properties of QALYs. Journal of Health Economics 16, 155-175.

Bleichrodt, H. and Johannesson M., 1997. The Validity of QALYs: An Empirical Test of Constant Proportional Tradeoff and Utility Independence. Medical Decision Making 17, 21-32.

Bleichrodt, H., Pinto J. L., and Wakker P. P., 2001. Using Descriptive Findings of Prospect Theory to Improve the Prescriptive Use of Expected Utility. Management Science 47, 1498-1514.

Bostic, R., Herrnstein R. J., and Luce R. D., 1990. The Effect on the Preference Reversal of Using Choice Indifferences. Journal of Economic Behavior and Organization 13, 193-212.

Chapman, G. B., 1998. Similarity and Reluctance to Trade. Journal of Behavioral Decision Making 11, 47-58.

Clarke, P., Wolstenholme J., and Johnstone K., 2003. Time Trade-Off of Time Gain: It All Depends on Your Point of Reference. Working Paper University of Oxford.

de Wit, G. A., van Busschbach J. J., and de Charro F. T., 2000. Sensitivity and Perspective in the Valuation of Health Status. Health Economics 9, 109-126.

Dolan, P., 1997. Modeling Valuations for EuroQol Health States. Medical Care 35, 1095-1108.

Dolan, P., 2000. The Measurement of Health-Related Quality of Life for Use in Resource Allocation Decisions in Health Care. In: Culyer, A. J. and Newhouse, J. P. (Eds.), Handbook of Health Economics, Vol. 1B. Elsevier Science, Amsterdam, 1723-1760.

Dolan, P., 2001. Personal communication.

Dolan, P., Gudex C., Kind P., and Williams A., 1996. Valuing Health States: A Comparison of Methods. Journal of Health Economics 15, 209-231.

Dolan, P. and Jones-Lee M., 1997. The Time Trade-Off: A Note on the Effect of Lifetime Reallocation of Consumption and Discounting. Journal of Health Economics 16, 731-739.

Dolan, P. and Stalmeier P. F. M., 2003. The Validity of Time Trade-Off Values in Calculating QALYs: Constant Proportional Time Trade-Off Versus the Proportional Heuristic. Journal of Health Economics (in press).

Hall, J., Gerard K., Salkeld G., and Richardson J., 1992. A Cost Utility Analysis of Mammography Screening in Australia. Social Science and Medicine 34, 993-1004.

Hershey, J. C. and Schoemaker P. J. H., 1985. Probability versus Certainty Equivalence Methods in Utility Measurement: Are They Equivalent? Management Science 31, 1213-1231.

Johannesson, M., Pliskin J. S., and Weinstein M. C., 1994. A Note on QALYs, Time Tradeoff, and Discounting. Medical Decision Making 14, 188-193.

Kahneman, D. and Tversky A., 1979. Prospect Theory: An Analysis of Decision Under Risk. Econometrica 47, 263-291.

McNeil, B. J., Weichselbaum R., and Pauker S. G., 1981. Tradeoffs Between Quality and Quantity of Life in Laryngeal Cancer. New England Journal of Medicine 305, 982-987.

Miyamoto, J. M. and Eraker S. A., 1985. Parameter Estimates for a QALY Utility Model. Medical Decision Making 5, 191-213.

Miyamoto, J. M. and Eraker S. A., 1988. A Multiplicative Model of the Utility of Survival Duration and Health Quality. Journal of Experimental Psychology: General 117, 3-20.

Ortona, G. and Scacciati F., 1992. New Experiments of the Endowment Effect. Journal of Economic Psychology 13, 277-296.

Pliskin, J. S., Shepard D. S., and Weinstein M. C., 1980. Utility Functions for Life Years and Health Status. Operations Research 28, 206-223.

Richardson, J., 1994. Cost Utility Analysis: What Should Be Measured? Social Science and Medicine 39, 7-22.

Rutten-van Mölken, M. P., Bakker C. H., van Doorslaer E. K. A., and van der Linden S., 1995. Methodological Issues of Patient Utility Measurement. Experience from Two Clinical Trials. Medical Care 33, 922-937.

Sackett, D. L. and Torrance G. W., 1978. The Utility of Different Health States as Perceived by the General Public. Journal of Chronic Disease 31, 697-704.

Spencer, A., 2003. The TTO Method and Procedural Invariance. Health Economics (in press).

Stalmeier, P. F. M., 2002. Discrepancies between Chained and Classic Utilities Induced by Anchoring with Occasional Adjustments. Medical Decision Making 22, 53-64.

Stalmeier, P. F. M., Bezembinder T. G. G., and Unic I. J., 1996. Proportional Heuristics in Time Trade Off and Conjoint Measurements. Medical Decision Making 16, 36-44.

Stalmeier, P. F. M., Wakker P. P., and Bezembinder T. G. G., 1997. Preference Reversals: Violations of Unidimensional Procedure Invariance. Journal of Experimental Psychology: Human Perception and Performance 23, 1196-1205.

Stiggelbout, A. M., Kiebert G. M., Kievit J., Leer J. W. H., Stoter G., and de Haes J. C. J. M., 1994. Utility Assessment in Cancer Patients: Adjustment of Time Tradeoff Scores for the Utility of Life Years and Comparison with Standard Gamble Scores. Medical Decision Making 14, 82-90.

Sutherland, H. U., Llewellyn-Thomas H., Boyd N. F., and Till J. E., 1982. Attitudes Toward Quality of Survival: The Concept of 'Maximal Endurable Time'. Medical Decision Making 2, 299-309.

Tversky, A. and Kahneman D., 1991. Loss Aversion in Riskless Choice: A Reference-Dependent Model. Quarterly Journal of Economics 56, 1039-1061.

Tversky, A., Sattath S., and Slovic P., 1988. Contingent Weighting in Judgment and Choice. Psychological Review 95, 371-384.

van Busschbach, J. J., 1994. The Validity of Qalys. Gouda Quint, Arnhem, The Netherlands (in Dutch).