

Nueva evidencia sobre viejos desafíos para los Años de Vida Ajustados por la Calidad: el máximo tiempo tolerable y la inversión de las preferencias

Jorge-Eduardo Martínez Pérez.

Departamento de Economía Aplicada. Universidad de Murcia.

José Luis Pinto Prades.

Departamento de Economía y Empresa, Universitat Pompeu Fabra

José María Abellán Perpiñán.

Departamento de Economía Aplicada. Universidad de Murcia.

Correspondencia:

Jorge-Eduardo Martínez Pérez. Departamento de Economía Aplicada. Facultad de Economía y Empresa. Universidad de Murcia. Campus de Espinardo. 30100. jorgemp@um.es

José Luis Pinto Prades. Departamento de Economía y Empresa, Universitat Pompeu Fabra. Trias Fargas, 25-27, Barcelona, 08005. jose.pinto@upf.es

José María Abellán Perpiñán. Departamento de Economía Aplicada. Facultad de Economía y Empresa. Universidad de Murcia. Campus de Espinardo. 30100. dionisos@um.es

Introduction

This paper concerns the descriptive validity of QALY utility models. QALYs are commonly described as a multiplicative model because the joint utility function can be decomposed into a product of factors that depend, respectively, on health status and duration. The simplest and most widely used multiplicative QALY model is the linear QALY model in which the utility function for duration is assumed to be linear. On the contrary, the more general class of multiplicative nonlinear QALY models allow for utility curvature.

Three major preference foundations have been given to multiplicative QALY models, namely, expected utility (EU), rank-dependent utility (RDU), and prospect theory (PT). Each of these theories imposes some specific conditions on the individual preference relation in order to be represented by QALYs. For example, under EU linear utility function for duration implies risk neutrality (Bleichrodt et al., 1997) which holds by virtue of the EU axiom of linearity in probability. On the contrary, under RDU since linearity in probability is no longer assumed the linear QALY model requires constant marginal utility for life years (Bleichrodt and Pinto, 2004) instead of risk neutrality. Empirical evidence rejects both risk neutrality (*e.g.* McNeil et al., 1978; Stiggelbout et al., 1994) and constant marginal utility (Bleichrodt and Pinto, 2004). An exception to previous EU and RDU characterizations is the new test of the linear QALY model presented by Doctor et al. (2004). These authors show that it is possible to test the linear QALY model by testing a single condition called *constant proportional coverage* which is valid under EU, RDU, and PT. Doctor et al. found considerable support for the linear QALY model.

In this paper we report new evidence that challenges the validity of multiplicative QALY models with independence on the specific preference foundation

given. Our findings concern three basic notions that are more primary than those specific assumptions commonly used to distinguish between existing characterizations. In consequence, our results suggest being cautious to judge the validity of multiplicative QALY models even under more general utility theories like RDU and PT. Moreover, preferences we observe affect the validity of possible relaxations of multiplicative QALY models in which utility curvature is allowed to vary as a function of health state.

First, we present a test of a preference condition typically regarded as a ‘structural’ assumption underlies all characterizations of multiplicative QALY models. This condition is currently called *monotonicity in duration*. Let (q, t) be a typical chronic health outcome where q denotes health status and t duration in years. Let \succeq be a preference relation meaning ‘at least as preferred as’. We say that \succeq satisfies *monotonicity in duration* if for all $(q_1, t_1), (q_1, t_2)$ with $t_2 > t_1$ then either $(q_1, t_2) \succ (q_1, t_1)$ or $(q_1, t_2) \prec (q_1, t_1)$. Hence, according to monotonicity in duration longer (shorter) durations are preferred to shorter (longer) durations for a given health state. A basic counterexample to monotonic preferences is the phenomenon of maximum endurable time (MET) first reported by Sutherland et al. (1982). This phenomenon suggests that for highly dysfunctional health states it is possible that longer durations are preferred to shorter durations up to some threshold or MET after which shorter durations are preferred to longer durations. In next section we describe how MET preferences falsify multiplicative QALY models regardless the utility theory in which they are based on. Although there are some previous studies which have reported findings troubling for the validity of monotonicity in duration (*e.g.* Dolan, 1996; Stalmeier et al., 2001) we think that the test we present here is the most pure test of monotonicity performed hitherto in the context of the health outcomes. Similarities and differences with previous related studies are discussed in last section of the paper.

Second, we find robust evidence on preference reversals entirely choice-based. Since it is commonly assumed in decision analysis that ‘choice’ is the “gold standard” for measuring preferences the type of preference reversals we observe is worrying for the validity of prescriptive decision analyses based on QALYs. To the best of our knowledge previous evidence on choice-based preference reversals is restricted to the experiment reported by Bleichrodt and Pinto (2002). However, we emphasize that our finding is new since all the outcomes used were riskless whereas Bleichrodt and Pinto used risky and riskless outcomes. Participants in our experiment performed two qualitative tasks. They were asked to rank a set of chronic health outcomes which only differed in duration, *i.e.*, the health status held constant through the ranking task. Participants were also asked to make several choices of type (q_1, t_1) vs (q_1, t_2) with $t_2 > t_1$. A preference reversal arises when people prefer a longer duration in one task and a shorter duration in the other task, *e.g.*, $(q_1, t_1) \prec (q_1, t_2)$ in direct choice but $(q_1, t_1) \succ (q_1, t_2)$ when the choice is inferred from the rank-order of the set of health outcomes constructed for health state q_1 .

Third, violations of transitivity were originally suggested as an explanation to preference reversals (*e.g.* Slovic and Lichtenstein, 1968; Lindman, 1971; Grether and Plott, 1979). Also intransitivities are commonly regarded as normatively undesirable (MacCrimmon, 1968). Hence, evidence on intransitivities among health outcomes limits the prescriptive applicability of QALYs. Since rankings force choices to be consistent while simple choices do not, we test violations of transitivity by inspection of intransitive cycles, *i.e.*, $(q_1, t_1) \succeq (q_1, t_2)$ and $(q_1, t_2) \succeq (q_1, t_3)$ but also $(q_1, t_1) \succeq (q_1, t_3)$, exhibited by respondents in pairwise choices. The traditional choice-matching preference reversal has been also explained by means of the prominence hypothesis (Tversky et al., 1988). According to this hypothesis people are more likely to prefer the

alternative that is superior on the most prominent attribute of the available options in choice than in matching. The traditional explanation for the prominence effect is based on the compatibility between the nature of the task (qualitative or quantitative) and the type of decision strategy evoked (Fischer and Hawkins, 1993). In this way, a qualitative task as a choice is compatible with a qualitative decision strategy such as “lexicographic ordering”. Therefore, as choice is a qualitative task and matching is a quantitative task it is expected that subjects pay attention to the most salient attribute in choice whereas in matching tasks they pay attention to all attributes more equally. However, the two methods used in this study (ranking and pairwise choices) are qualitative tasks. Hence there is no reason a priori to assume that preferences for the salient attribute, *i.e.*, duration in our case, are more likely in choices than in rankings. We compare percent rates of preferences for the outcome with superior duration between tasks.

The paper is structured as follows. Section 2 briefly describes how violations of monotonicity falsify multiplicative QALY models regardless preference foundations. Next, in section 3 we explain the experiment conducted to test monotonicity, potential preference reversals across tasks, and violations of transitivity. Results are provided in section 4. Discussion closes the paper.

2. Violations of monotonicity falsifies multiplicative QALY models

We describe how violations of monotonicity contradict standard gamble invariance. Finally, in addition to standard gamble invariance, there are two key assumptions that the two non-linear QALY models more widely applied, *power* and *exponential models*, must obey with independence of the utility theory assumed. Whereas the power QALY model must satisfy *constant proportional risk posture*, the

exponential QALY model must satisfy *constant absolute risk posture* (e.g., Miyamoto and Eraker, 1989; Cher et al., 1997; Miyamoto, 1999; Bleichrodt and Miyamoto, 2003). Again, violations of monotonicity are showed that cause violations of power and exponential models.

- *The linear QALY model*

As noted in introduction, Doctor et al. (2003) have provided a new test of the linear QALY model. This characterization is valid if the individual satisfies either EU or RDU or PT. Next, we discuss how violations of monotonicity with respect to life years imply that the new test of Doctor et al. must be also violated. The central condition described by Doctor et al., *constant proportional coverage*, was first derived by Miyamoto (1999) in order to characterize the linear QALY model under RDU axioms (Definition 14, p. 227; theorem 16, p. 229). This condition is implied by risk neutrality, so it is also valid under EU. Doctor et al. show that constant proportional coverage must also hold under PT assumptions.

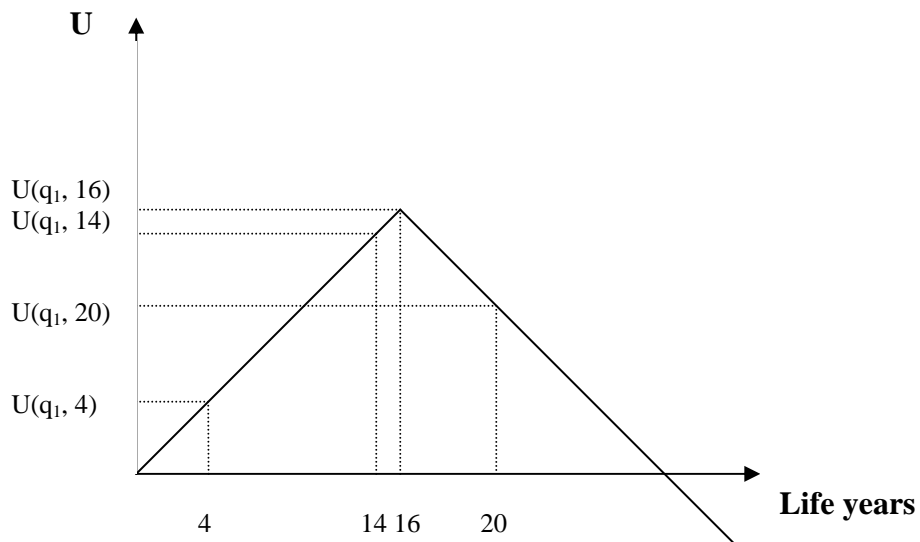
To define the constant proportional coverage assumption we need to introduce some notation and basic properties. Let $[(q_1, t_1), p; (q_2, t_2)]$ be a typical lottery where chronic health outcome (q_1, t_1) results with probability p . Let \succeq be the preference relation ‘at least as preferred as’ defined over lotteries. As usual, strict preference is denoted by \succ and indifference by \sim . By setting $p = 1$ in $[(q_1, t_1), p; (q_2, t_2)]$, \succeq defines a preference relation over chronic health outcomes. If $(q, t_2) \sim [(q, t_1), p; (q, t_3)]$ for any q and t_1, t_2, t_3 , then p is called the *probability equivalent* of (q, t_2) with respect to the end points (q, t_1) and (q, t_3) . In the same way, t_2 will be called the *certainty equivalent* of the lottery with respect to the health state q .

Consider any health state q and durations $t_1 > t_2 > t_3$ and $t'_1 > t'_2 > t'_3$, we say that *constant proportional coverage* (CPC) holds if $(q, t_2) \sim [(q, t_1), r; (q, t_3)]$, $(q, t'_2) \sim [(q, t'_1), s; (q, t'_3)]$ and $(t_2 - t_3)/(t_1 - t_3) = (t'_2 - t'_3)/(t'_1 - t'_3)$, then $r = s$.

The rationale of this assumption is that under the linear QALY model probability equivalents r and s are a function of ratios $(t_2 - t_3)/(t_1 - t_3)$ and $(t'_2 - t'_3)/(t'_1 - t'_3)$ respectively (see Miyamoto, 1999, p. 226, for an example with 50/50 lotteries). In consequence certainty equivalents t_2 and t'_2 cover a constant proportion of the $[t_1, t_3]$ and $[t'_1, t'_3]$ ranges.

Suppose now that preferences are non-monotonic in duration. Specifically, in order to illustrate the relationship between violations of monotonicity and violations of CPC suppose that, for a given health state q_1 , utility function for duration behaves according to the MET phenomenon. This case is depicted in Figure 1.

Figure 1. Non-monotonicity and the linear multiplicative QALY model



We know that if an individual is asked to determine the probability r that yields indifference between $[(q_1, 20), r, (q_1, 4)]$ and $(q_1, 14)$, then, by CPC, that probability s that ensures indifference between $[(q_1, 15), s, (q_1, 7)]$ and $(q_1, 12)$ should be identical to the first elicited probability equivalent r . Note that durations have been chosen such that ratios $(t_2 - t_3)/(t_1 - t_3) = (t'_2 - t'_3)/(t'_1 - t'_3) = 5/8$. However, as in our example the utility of 14 years in health state q_1 is higher than utilities of 4 years and 20 years respectively, a duration of 14 years cannot be a certainty equivalent of $[(q_1, 20), r, (q_1, 4)]$. In other words, 14 years in health state q_1 will be always preferred to a lottery which ranges from 4 to 20 years because preferences are not monotonically increasing. The extreme case would be if the MET was reached for 14 years (and not 16 years as it is represented in figure 1). Then 14 years would not ever be a certainty equivalent because it would be now the absolute maximum for all possible durations.

- *The general multiplicative QALY model*

Miyamoto (1999) derived a general multiplicative QALY model, *i.e.*, without assuming a specific functional form for the utility function over duration, by using a single condition valid under both EU and RDU. This preference condition claims that certainty equivalents to binary gambles with $p=0.5$ are invariant under changes in the health status. A more general formulation of this assumption is given in Miyamoto et al. (1998) with the name of *standard gamble invariance* (SGI). This condition asserts that for two health states q_1 and q_2 unequal to death, if $(q_1, t_2) \sim [(q_1, t_1), r; (q_1, t_3)]$ then $(q_2, t_2) \sim [(q_2, t_1), r; (q_2, t_3)]$.

SGI is similar to CPC in the sense that imposes equality between probability equivalents. That is, SGI says that if r is the probability equivalent of (q_1, t_2) with respect to (q_1, t_1) and (q_1, t_3) , then r is also the probability equivalent of (q_2, t_2) with

respect to (q_2, t_1) and (q_2, t_3) . The difference is that whereas CPC requires that probability equivalents are the same in different durations, SG invariance requires that probability equivalents are the same in different health states.

Figure 2. Non-monotonicity and the general multiplicative QALY model

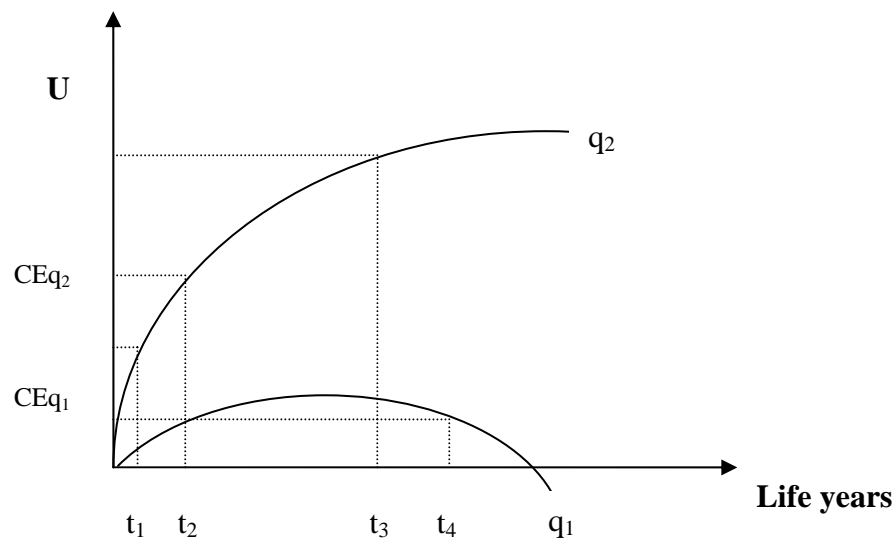


Figure 2 illustrates how a violation of monotonicity falsifies the general multiplicative QALY model¹. Suppose that t_2 is the certainty equivalent of a lottery where the health status is fixed at q_1 . As figure 2 shows the utility function for duration in that health state increases monotonically. SGI requires now that the certainty equivalent remains constant if we replace the health state q_1 by q_2 . However, as utility function for health state q_2 is non-monotonic in duration we can see that the certainty equivalent is not necessarily unique. If the individual set t_4 as the certainty equivalent of the lottery in health state q_2 we have then two different certainty equivalents. A violation of SGI follows. Other different counterexample to the general multiplicative

¹ As Miyamoto et al. (1998) noted, non-monotonicity in a specific health state falsifies the general multiplicative QALY model only if there are other health states whose utilities are not extreme at the same value. That is, only if utility is strictly increasing or decreasing in other health states.

QALY model was discussed by Miyamoto et al. (1998: p. 845) within the realm of EU. It is also valid for RDU.

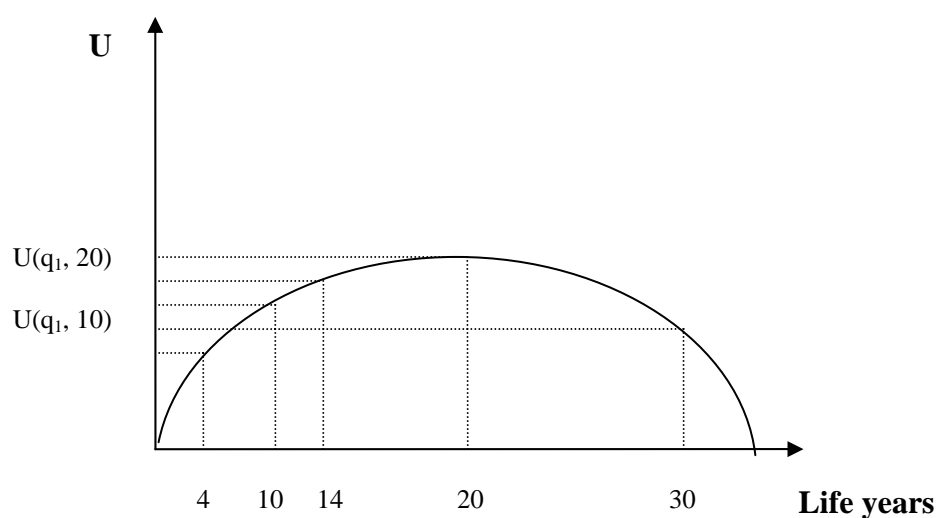
- *Power and exponential multiplicative QALY models*

Power and exponential utility functions for durations are characterized by *invariants* of the preference order (Pratt, 1964). Specifically, in the health domain, power utility for duration can be obtained by assuming that, for a given health state, a preference between lotteries does not vary if all durations involved are multiplied by the same constant duration (*i.e.*, constant proportional risk attitude). Instead, exponential utility requires that preference between lotteries remains invariant if a constant duration is added to all durations involved (*i.e.*, constant absolute risk attitude).

Failures of monotonicity with respect to life duration violate power and exponential QALY models. Next, we describe an example for the power case. Exponential forms can be contradicted in a similar way.

According to the example displayed in figure 3, we can find that a lottery with endpoints $(q, 10)$ and $(q, 2)$ is preferred to certain outcome $(q, 5)$. Similarly, suppose that a lottery with endpoints $(q, 15)$ and $(q, 7)$ is also preferred to $(q, 10)$. Imagine that now we multiply each duration involved by a constant duration $c = 2$. Constant proportional risk attitude requires that the new lottery with endpoints $(q, 10 \times 2 = 20)$ and $(q, 4)$ remains regarded as preferred to the new outcome $(q, 10)$, and that the lottery with endpoints $(q, 15 \times 2 = 30)$ and $(q, 14)$ is also preferred to $(q, 20)$. However, as the MET is reached for 20 years we can observe that a preference reversal would occur. That is, $[(q, 20), r; (q, 4)] \succ (q, 10)$ but $[(q, 30), r; (q, 14)] \prec (q, 20)$. This obviously falsifies the power QALY model.

Figure 3. Non-monotonicity and the multiplicative power QALY model



4. Experiment

- Participants

Participants were 90 economics students from the University of Murcia. The experiment was carried out in small group sessions with at most five subjects per group. Each participant attended three experimental sessions, one to rank-order health outcomes (henceforth, the ranking session) and other two to choose between health outcomes (choice sessions). In order to avoid order and memory effects, tasks were randomly assigned to participants and sessions were separated by one week each.

-Health outcomes

We used seven EQ-5D health states. Table 1 shows the description of the health states. Throughout the experiment health states were anonymously labelled P-W.

Table 1. The description of the EQ-5D health states

STATE T	
1	No problems walking about
1	No problems with self care
1	No problems with performing usual activities
1	No pain or discomfort
2	Moderately anxious or depressed
STATE U	
1	No problems walking about
1	No problems with self care
1	No problems with performing usual activities
1	No pain or discomfort
3	Extremely anxious or depressed
STATE V	
1	No problems walking about
1	No problems with self care
3	Unable to perform usual activities
1	No pain or discomfort
2	Moderately anxious or depressed
STATE W	
1	No problems walking about
2	Some problems washing or dressing self
2	Some problems with performing usual activities
2	Moderate pain or discomfort
3	Extremely anxious or depressed
STATE X	
1	No problems walking about
3	Unable to wash or dress self
3	Unable to perform usual activities
3	Extreme pain or discomfort
2	Moderately anxious or depressed
STATE Y	
3	Confined to bed
3	Unable to wash or dress self
2	Some problems with performing usual activities
3	Extreme pain or discomfort
2	Moderately anxious or depressed
STATE Z	
3	Confined to bed
3	Unable to wash or dress self
3	Unable to perform usual activities
3	Extreme pain or discomfort
3	Extremely anxious or depressed

The health states were chosen with a view to cover the range of tariff values of the EuroQol system. We follow this criterion as a check on the consistency of responses with the EuroQol algorithm. It would be logical to expect that the proportion of

participants who prefer shorter durations to longer durations increases as the health state gets worse. Table 2 shows the tariff values of the health states selected according to the EuroQol algorithm for the case in which coefficients of the algorithm were obtained from time tradeoff (TTO) valuations.

Table 2. Tariff values for the EQ-5D health states

EQ-5D health states	Social tariffs (TTO values)
11112	0.9095
11113	0.5388
11312	0.4223
12223	0.2442
13332	-0.1451
33232	-0.4439
33333	-0.6533

From the combination of each health state with durations 0, 13, 24, 38, and 57 years respectively, we obtained the five health outcomes presented to participants. We set the maximum duration in 57 years in order to do not exceed the life-expectancy of participants (the mean age was about twenty years).

- Tasks

Before the first experimental session participants were introduced to the experiment receiving some notions on the EuroQol system and how impaired conditions are described by this system. Also, before each session participants were instructed to make choices or performing rankings according their true preferences, even though it implied that less years were preferred to more years in a given health state. This type of decreasing preferences was described by some hypothetical examples. Questionnaires began with a trial question that was checked with participants before experimental questions were answered.

In the ranking session, participants were asked to rank-order the five health outcomes for each of the seven EQ-5D health states from more to less preferred. Outcomes were printed on a set of cards which were distributed in a random order. To avoid response errors, participants were asked to confirm their rankings. If they did not confirm, they could change the ordering. We repeated the process until participants did agree the orderings revealed.

In choice sessions, participants were asked to make choices between two health outcomes. We constructed ten pairs of health outcomes for each EQ-5D health state. For example, for health state X we presented the following comparisons: (X, 57 yrs.) vs (X, 38), (X, 57) vs (X, 24), (X, 57) vs (X, 13), (X, 57) vs (X, 0), (X, 38) vs (X, 24), (X 38) vs (X 13), (X 38) vs (X, 0), (X, 24) vs (X, 13), (X, 24) vs (X, 0), (X, 13) vs (X, 0). Overall, each participant made seventy choices, *i.e.*, 10 pairs \times 7 health states, distributed across two questionnaires including 35 choices each. The order in which choices were presented was random. Each questionnaire was administered in a different experimental session. To avoid response errors, participants were asked to confirm their choices by filling in a table. This additional task forced them to check earlier responses.

- *Methods*

To test monotonicity and transitivity at the individual level we classified each participant into one of seven possible preference patterns:

- i. Individuals whose preferences are uniformly increasing monotonic, *i.e.*, they have increasing monotonic preferences for a given task in each health state.
- ii. Individuals whose preferences are uniformly decreasing monotonic.
- iii. Individuals whose preferences are uniformly non-monotonic.
- iv. Individuals whose preferences are uniformly intransitive.

- v. Individuals with a mixed pattern of monotonic preferences, *i.e.* they have increasing monotonic preferences for some health states and decreasing monotonic preferences for others health states.
- vi. Individuals with a mixed pattern of non-monotonic preferences, *i.e.*, they have non-monotonic preferences for at least one health state but not for all health states.
- vii. Individuals with a mixed pattern of intransitive preferences, *i.e.*, they have intransitive preferences for at least one health state but not for all health states.

To test monotonicity at the aggregate level we calculated for each health state and task both the percent rate $P(m)$ of participants for whom preferences were monotonic and the percent rate $P(\text{non-m})$ of participants with non-monotonic preferences. If each participant who prefers (q_1, t_1) to (q_1, t_2) , with $t_2 > (<) t_1$, also prefers (q_1, t_3) to (q_1, t_4) , with $t_3 > (<) t_4$, and so on, then $P(m) > P(\text{non-m})$ given health state q_1 that is the hypothesis we test in aggregate data. Those participants who exhibited intransitive preferences were excluded from the test of monotonicity because they yielded cyclical rankings. We also compared the percent rate $P(t)$ of participants for whom preferences were transitive with the percent rate $P(i)$ of participants with intransitive preferences. Transitivity requires that $P(t) > P(i)$ for each health state.

Both monotonicity and transitivity were tested by using the *goodness-of-fit* Chi-squared (χ^2) test. We also tested whether the probability of exhibiting non-monotonic preferences depended on the task by using the nonparametric McNemar test and/or if they depended on the health status by the nonparametric Cochran Q test.

Since the falsification of the general multiplicative QALY model requires that utility in different health states is not extreme at the same value of the MET (Miyamoto et al., 1998) we tested whether the extreme value corresponding to the modal preference

ordering, *i.e.*, the most frequent ranking for each health state, differed across the seven EQ-5D health states. For example, suppose that for health states T-X 57 years is the most frequent preferred duration and that for states Y and Z the same duration is the most frequent dispreferred duration. We would then test if the probability of being 57 years the most/least preferred duration is independent on the health state. Significance of differences across the seven health states was tested by the nonparametric Cochran Q test.

To test the existence of preference reversals we calculated the percent rate of preference reversals for each health state as the fraction of people who switch their choices from the first task to the second task. In order to check whether preference reversals were important regardless intransitivities we computed the rates both with and without participants who yielded any intransitivity. In addition, as a way of testing whether lexicographic ordering could explain preference reversals we calculated the percent rate of “prominent preferences” for the two tasks. This rate was calculated as the fraction of people who preferred the outcome that was superior with respect to the most salient attribute, *i.e.*, duration. This hypothesis was tested by the nonparametric McNemar test. Intransitive respondents were excluded from the analysis.

5. Results

-Monotonicity and transitivity

At the individual level, we find that most participants display a mixed pattern of non-monotonic preferences. Thirty-seven participants behave consistently according this pattern in the choice task and sixty-two in the ranking task. Therefore, 41.1% out of participants violate monotonicity in at least one health state with choices whereas 68.89% violate monotonicity in at least one health state with rankings. No participant is uniformly intransitive but twenty participants make sometimes intransitive choices.

Eighteen out of these subjects with a mixed pattern of intransitive preferences in the choice task exhibit a mixed pattern of non-monotonic preferences in the ranking task. After removing these participants from ranking data the percent rate of mixed non-monotonic preferences is very similar, *i.e.*, 62.86%. Twenty-seven participants display a mixed pattern of monotonic preferences with choices and twenty-six with rankings. We find that six individuals are uniformly increasing monotonic in the choice task and only one in the ranking task. No participant is uniformly decreasing monotonic or uniformly non-monotonic in the choice task. No participant is uniformly decreasing monotonic in the ranking task.

In figures 4 and 5 the percentage of participants with monotonic preferences is displayed by the height of grey columns. Numerical percentages are also showed. Black and dotted bars display percentages of participants with nonmonotonic preferences and intransitive preferences respectively.

Figure 4. Choice task: percent rate of monotonic, non-monotonic, and intransitive preferences (N = 90)

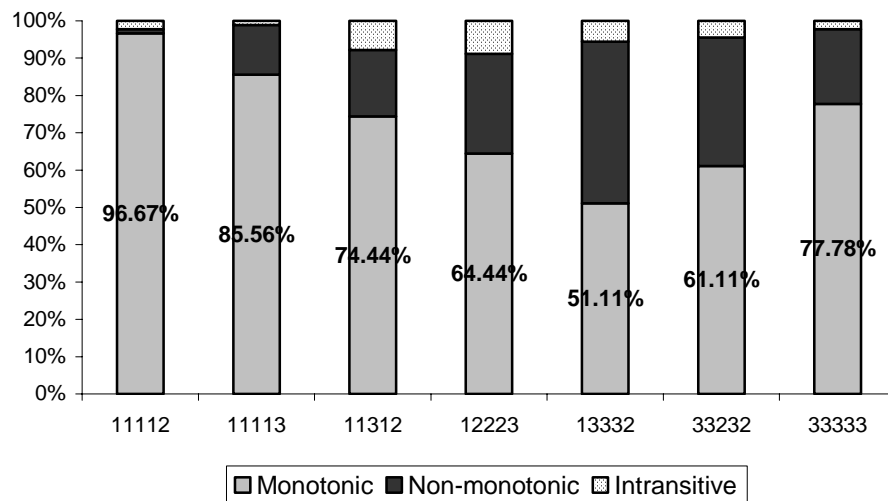
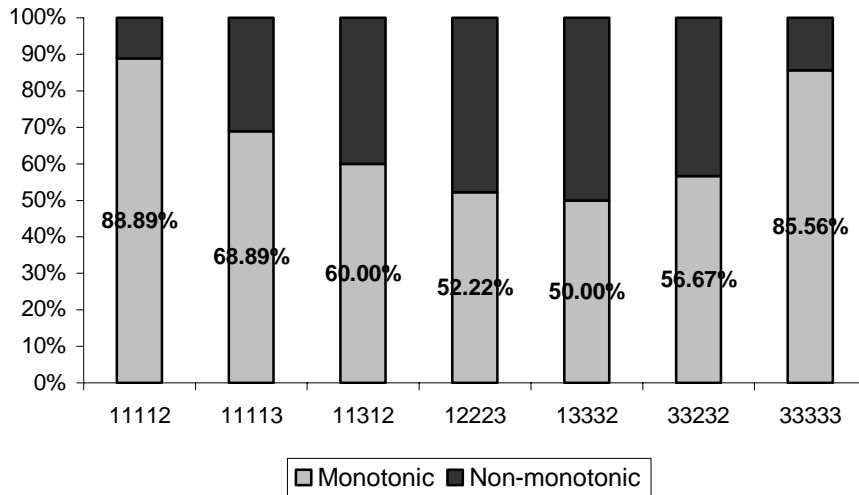


Figure 5. Ranking task: percent rate of monotonic and non-monotonic preferences (N = 90)



Four main points arise from the inspection of figures 4 and 5. First, percent rates of non-monotonic preferences range from 1.11% (state 11112) to 43.33% (states 13332) under the choice task, and from 11.11% to 50% (for the same states) under the ranking task. Second, percent rate of monotonic (non-monotonic) preferences decreases (increases) with severity level from health state 11112 to state 13332, for which the lowest (highest) rate is reached. Third, we observe that percentages of monotonic preferences are higher for simple choices than for rankings. In consequence, our data suggest that non-monotonic preferences are more likely to happen when the subject is asked to perform preference rank-orderings rather than making pairwise choices. Four, percent rates of intransitivities are relatively small. It ranges from 1.11% for health state 11113 to 8.89% for health state 12223. We obtain significant differences between percent rates of transitive and intransitive preferences supporting transitivity in all health states (chi-square, $P < 0.0001$).

Figures 6 and 7 show percentages of monotonic and non-monotonic preferences after those participants with intransitive preferences are excluded. Bars with horizontal straight lines denote percent rates of increasing monotonic preferences. Bars with vertical straight lines denote percent rates of decreasing monotonic preferences. Black

bars denote non-monotonic preferences. Stars indicate statistical significance at $\alpha = 0.001$.

Figure 6. Choice task: percent rate of monotonic (increasing/decreasing) and non-monotonic preferences (N = 70)

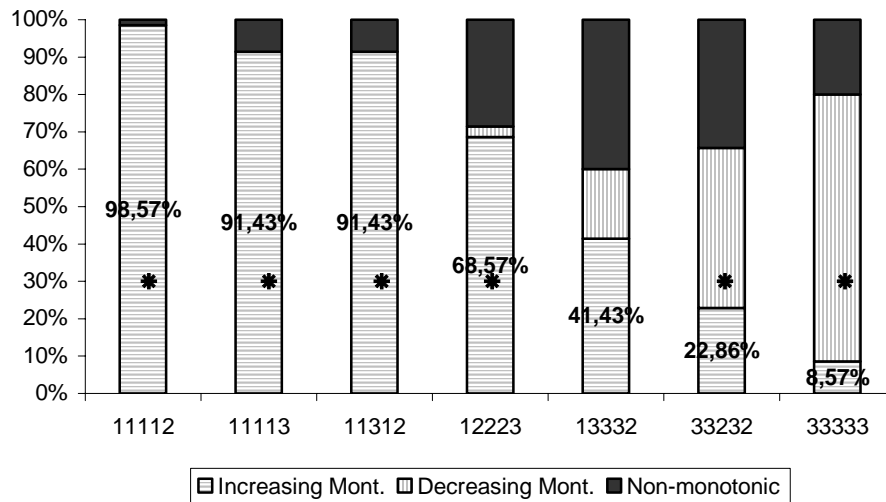
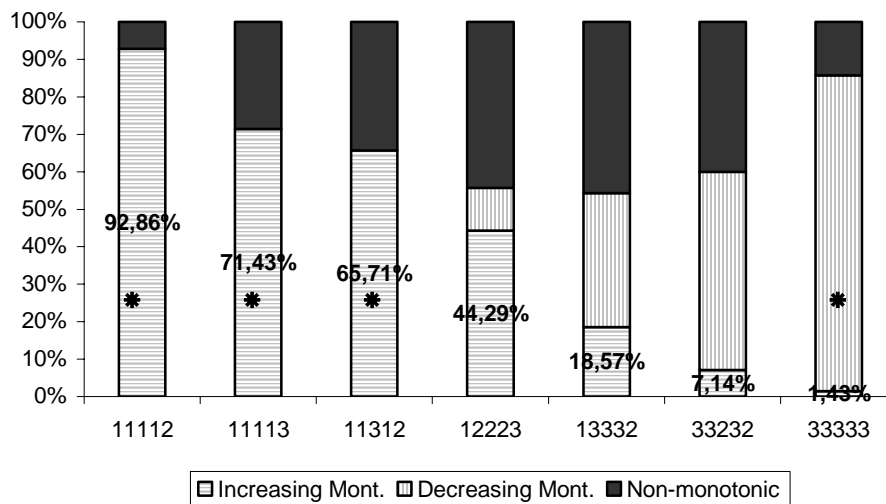


Figure 7. Ranking task: percent rate of monotonic (increasing/decreasing) and non-monotonic preferences (N = 70)



It can be seen that increasing preferences decrease with severity in a consistent way with EuroQol tariffs. As EQ-5D health state utilities get more negative, the fraction of decreasing preferences increases, being the rule for health states 33232 and 33333.

Under the choice task (figure 6) we observe that the rate of monotonic preferences is significantly higher than the rate of non-monotonic preferences in all cases except for health state 13332 (Chi-square, $P = 0.094$). We note that although discrepancies between monotonic and non-monotonic percent rates are statistically significant in the direction predicted by monotonicity, there are important rates of non-monotonic preferences for health states 12223 and 33232, *i.e.*, 28.57% and 34.29% respectively.

Figure 7 (ranking task) shows more robust evidence contrary to monotonicity in duration at the aggregate level. In particular, we do not find significant differences between monotonic and non-monotonic rates for health states 12223, 13332, and 33232 (chi-square, $P = 0.339$, $P = 0.473$, and $P = 0.094$, respectively). Percent rates of non-monotonic preferences for these states are 45.71%, 44.29%, and 44% respectively. Percentages are also high for health states 11113 and 11312, *i.e.*, 28.57% and 34.29% respectively, although monotonicity cannot be rejected.

It is apparent that monotonicity is more frequently violated with rankings than with pairwise choices. Indeed we find that the probability of exhibiting non-monotonic preferences is significantly higher in ranking than in choice for health states 11113, 11312, and 12223 by the McNemar test ($P < 0.001$ in the two first cases; $P < 0.05$ in the third case). In addition, it seems that the probability of non-monotonic preferences arises is not independent on the health status (Cochran Q test, $P < 0.0001$ for both ranking and choice tasks). As it can be seen, percent rate of non-monotonic preferences increases with severity level from health state 11112 to state 13332, for which the highest rate is reached.

We find that for all health states 57 years is the most frequent extreme duration, *i.e.*, maximum/minimum. We then test whether the probability of 57 years is the most

frequent extreme duration is independent on the health status. Independence on the health state is rejected by the Cochran Q test ($P < 0.0001$ for both ranking and choice tasks). This means that although 57 years is the modal extreme value for all health states there exist significant variation between health states as a consequence of non-monotonicity. This finding suggests that non-monotonic preferences indeed challenge the validity of multiplicative QALY models.

- Preference reversals

The proportion of preference reversals was 1.51%, 18.97%, 24.87%, 33.02%, 21.92%, 13.54%, and 6.16% for health states 11112, 11113, 11312, 12223, 13332, 33232, and 33333 respectively. Intransitivities only explain these reversals in a small part. On average, less than 5 percent of reversals reflect intransitivity.

We then tested if reversals in preference were due to more salience of attribute duration in one task than in the other one. Figures 8-14 show percent rates of preferences for the health outcome superior in duration. Intransitive respondents were excluded from data. Stars indicate statistical significance at $\alpha = 0.01$. Crosses indicate statistical significance at $\alpha = 0.05$.

We observe that preferences for the outcome superior in duration are the rule for health states 11112, 11113, 11312, and 12223 under the choice task. Also, they are frequent for health state 13332. We find significant differences between choices and rankings for health states 11113, 11312, and 12223 and all pairs of durations except for the comparison 13 years vs 0 years. Significant differences are also found for comparisons 57 vs 38, 57 vs 24, 57 vs 13, 38 vs 24, and 24 vs 13 with health state

13332. One significant difference is found for comparison 57 vs 38 with health state 33232.

Figure 8. Health state 11112: percent of preferences for the outcome that has higher duration

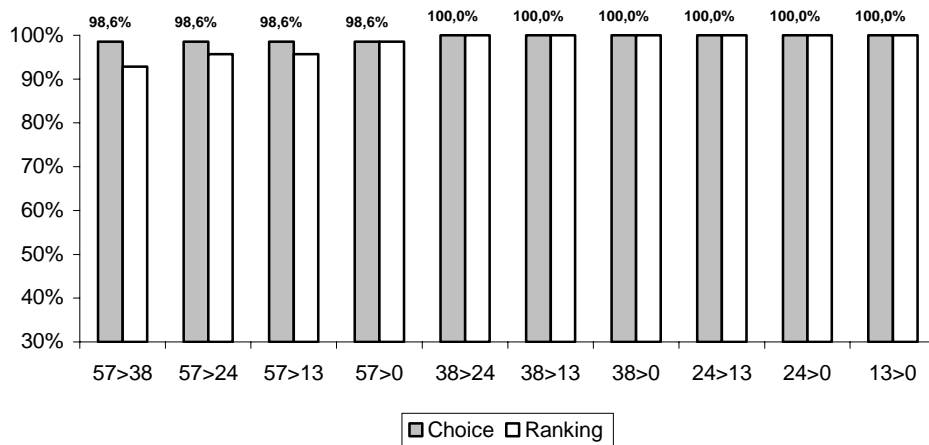


Figure 9. Health state 11113: percent of preferences for the outcome that has higher duration

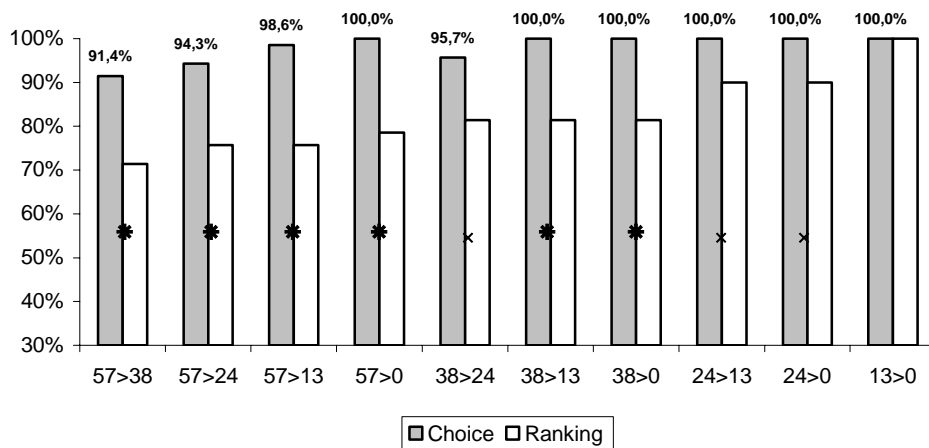


Figure 10. Health state 11312: percent of preferences for the outcome that has higher duration

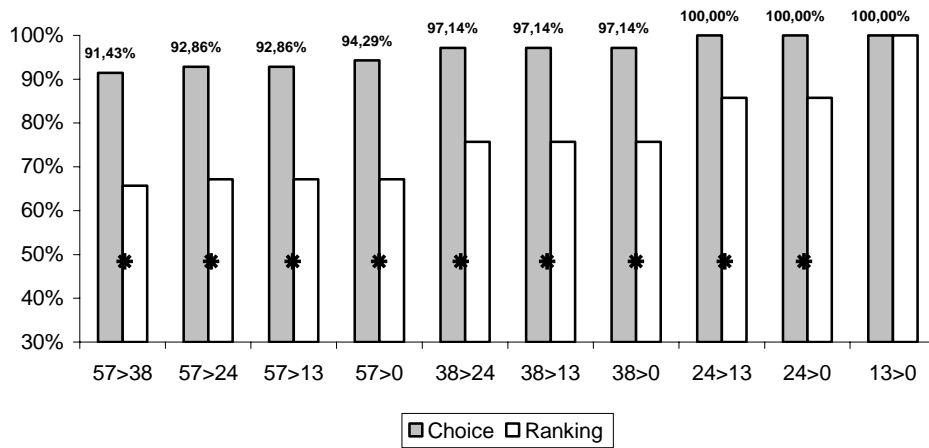


Figure 11. Health state 12223: percent of preferences for the outcome that has higher duration

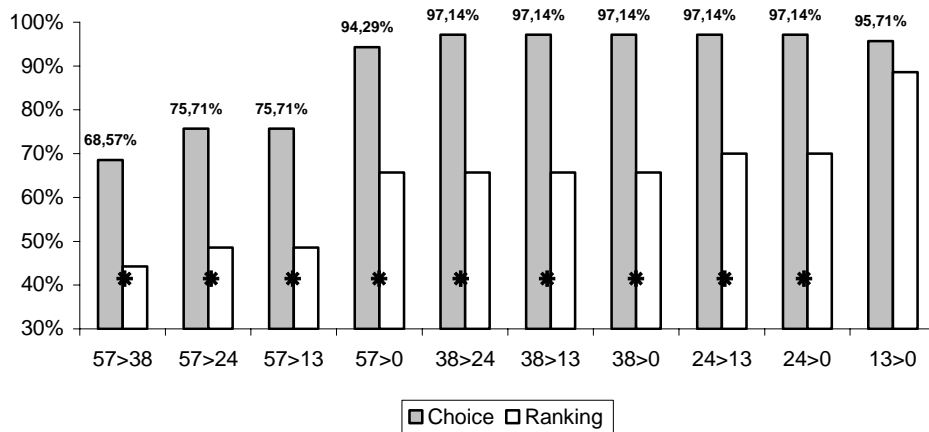


Figure 12. Health state 13332: percent of preferences for the outcome that has higher duration

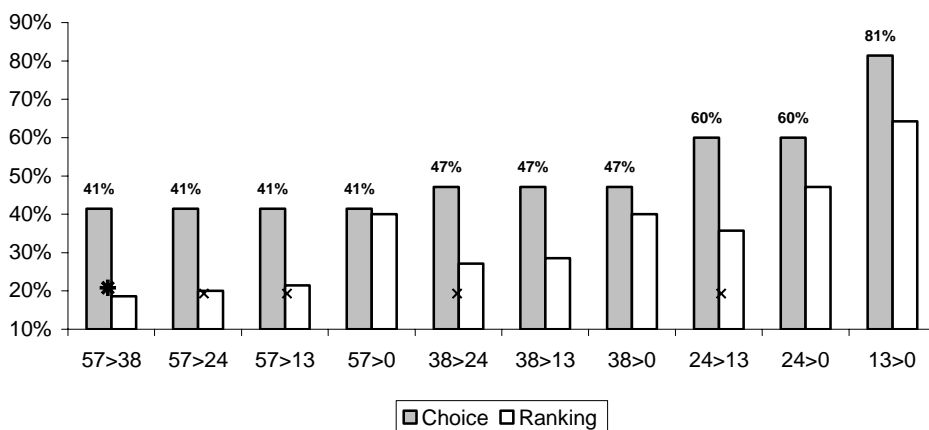


Figure 13. Health state 33232: percent of preferences for the outcome that has higher duration

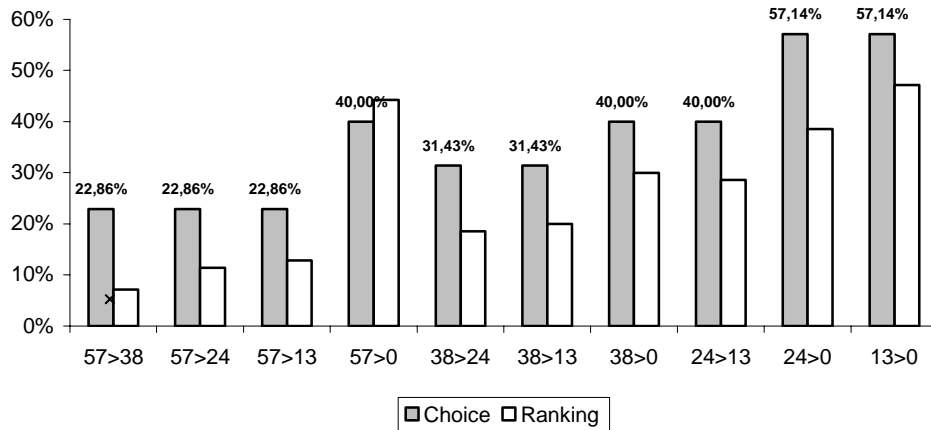
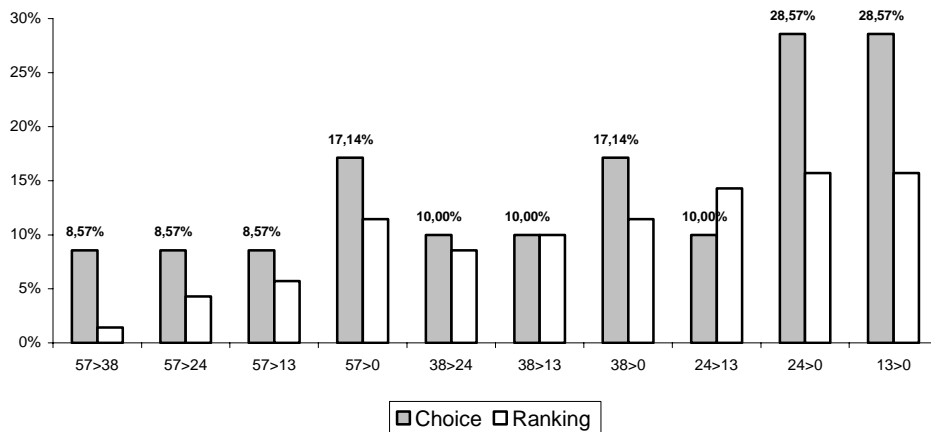


Figure 14. Health state 33333: percent of preferences for the outcome that has higher duration



6. Discussion

- Main findings

We find violations of three basic assumptions underlie QALY utility models. We find that monotonicity with respect to life years is frequently violated across the seven EQ-5D health states selected for this study. We use two different procedures to elicit preferences, namely: pairwise choices and rankings. At the individual level

violations range from around 41% with choices to almost 69% with rankings. At the aggregate level, the mean rate of violations is around 19% with choices and around 32% with rankings. For some health states the percent rate of non-monotonic preferences is around 45%. We observe that violations of monotonicity increases with severity of health status although the maximum rate of violations is not obtained for the most severe health state, *i.e.*, 33333, but for health state 13332.

We find new evidence on preference reversals with two elicitation procedures choice-based. Percent rates of preference reversals range from 1.51% for health state 11112 to 33.02% for state 12223. We find considerable support for the hypothesis that the preference reversals presented in this study can be explained by greater salience of the duration attribute in the choice task.

Finally, we also find some evidence on violations of transitivity. However, the intransitivities we observe are scarce (the highest rate is 8.89% for health state 12223) and it can only explain partly the reversals in preference.

- Previous related studies

Axiomatic definition of monotonicity in duration requires to compare several different durations for a given health state by means of choice-based procedures. Moreover, as Miyamoto et al. (1998) note, non-monotonicity in a specific health state falsifies multiplicative nonlinear QALY models only if there are other health states whose utilities are not extreme at the same value. Our design allows us to check this question. None previous study has adopted this approach to test monotonicity.

Sutherland et al. (1982) found that the proportion of respondents preferring death increased as duration increased, which reveals only partly that preferences are non-monotonic with respect to life years because all comparisons are performed with

respect to $(q, 0)$. For example, we find non-monotonic preference patterns such that any pair (q, t) is preferred to death but less years in q are preferred to more years in q , *e.g.*, for ten participants find that $(13332, 13 \text{ yrs.}) \succ (13332, 24) \succ (13332, 38) \succ (13332, 57) \succ (13332, 0)$.

Gudex and Dolan (1995) found that median visual analogue scale (VAS) scores for poor health states lasted 10 years were lower than when the same states lasted 1 year or 1 month. Dolan (1996) used the same durations (1 month, 1 year, and 10 years) to estimate new tariffs for all EQ-5D health states based on the VAS valuations obtained for a selection of states. This design displays violations of monotonicity in an indirect way, because it does not use choices but ratings. Moreover, neither Gudex and Dolan (1995) nor Dolan (1996) checked whether the implied ranking from the rating exercise agreed fully with direct ranking.

Finally, in various studies conducted by Stalmeier and colleagues different durations have been compared in the same health state. However, given that they were mainly interested in testing inconsistencies between direct choices and TTO values these studies only used two durations in direct choices. For example, participants in the experiment conducted by Stalmeier et al. (1996) were asked whether or not prefer living 25 years with metastasized breast cancer to living 50 years with the same health state. It is apparent that a single question involving only two durations is a test of monotonicity quite weak. Other limitation is that they only used one health state.

We find percent rates for health state 13332, *i.e.*, 40% and 44.29% for choices and rankings respectively, close to that reported by Stalmeier et al. (2003), *i.e.*, 48%, for EQ-5D state 21223. Our percentages are also similar to that found by Stalmeier et al. (2001), *i.e.*, 44%, for ‘skeletal metastases’ health state. If decreasing monotonic

preferences in our experiment are treated as non-monotonic preferences under the assumption that may exist some duration between 0 and 13 years for which more life years are preferred to less life years then we obtain rates of non-monotonic preferences for health states 12223, 13332, 33232, and 33333 higher than those reported by all previous studies (Stalmeier et al., 1996; 1997; 2001; 2003) that have used (as we do) choices to test monotonicity in preferences.

Bleichrodt and Pinto (2002) report evidence on preference reversals entirely choice-based. Rates of preference reversals are similar in the two studies. However, Bleichrodt and Pinto used riskless and risky health outcomes whereas we use riskless outcomes only. In consequence, the argument used by these authors to explain their preference reversals, *i.e.*, that people use different utility functions to evaluate riskless and risky outcomes, cannot explain our preference reversals. We find that our reversals can be explained by greater salience of attribute duration in the choice task. In some respect, this finding resembles the conclusion of the qualitative investigation of Robinson et al. (1997). These authors investigated the reason why 83.7% of respondent in the UK survey conducted by Dolan (1995) to estimate tariff values of the EuroQol rated at least one health state worse than death in the time trade-off (TTO), while rating it better than dead in the VAS. Robinson et al. suggest that this reversal is due to respondents ignore the duration of the health state when completing the VAS, being the same duration of the same health state more salient in the TTO. In a similar way, our results suggest that even though the two tasks we use, choices and rankings, may evoke the same decision strategy, *i.e.*, lexicographic ordering, it seems that duration is more salient in choices than in rankings leading to an overvaluation of duration respect to health status. On the contrary, it seems that participants weight both attributes more equally in rankings leading to less preference for outcomes with higher duration. This

explanation is also similar to hypotheses proposed to explain the prominence effect (Fischer et al., 1999). The novelty, however, is that salience or prominence of attributes can explain preference reversals between qualitative decision tasks.

Our evidence on intransitive cycles is scarce. Percent rates of intransitivities are lower than those reported by Tversky et al. (1990) and Cox and Grether (1991) to analyze the traditional choice-matching discrepancy. The difference is even higher with respect to other studies that used a pure-choice design similar to ours (*e.g.* Loomes et al., 1991). However, all these studies used choices between risky and riskless outcomes, all of them defined as monetary payments.

- Limitations

This study is not exempted from limitations. First, participants in our sample were young students. This might imply that negative adaptation to a hypothetical poor health state was higher than adaptation of less healthy people. Nevertheless, as we noted above proportions of non-monotonic preferences were similar to those reported in other studies. Other objection may concern the sample-size used in our analysis. Although it is larger than samples used in most previous studies (Gudex and Dolan, 1995; Stalmeier et al., 1996; 2001; Dolan and Stalmeier, 2003) it would be interesting to replicate our tests with a larger sample. Participants in our experiment did not receive financial compensation. Instead, experimental sessions were computed as half a credit. It would be interesting to check if results are robust to changes in compensation. However, we do not believe that financial motivation may vary our findings as it is supported by some studies (*e.g.*, Mellers et al., 1992). Indifferences between outcomes were not allowed. Hence, some choices might be forced and this might yield random error. However, with random choices one would expect a 50% rate of non-monotonic preferences for mild and severe health states alike. On the contrary, we find that violations of monotonicity

depend on the severity of the health status. A last objection could be that our choices were too simple, inducing easily salience-based decision. It is interesting to note that prominence effect seems to be robust even when each pair of choice alternatives is designed to be perceived as roughly equally attractive. With this type of “balanced” design Fischer et al. (1999) found that 63 percent of choices were in the direction of the prominence effect. In addition, as we noted in section 5, before each session participants were told that any type of choice and ranking was allowed. Various examples illustrating hypothetical choices and rankings of type “I prefer less years to more years in this health state” were described, and one trial question was performed and checked with participants.

- Implications

From our study can be inferred that the maximum endurable time phenomenon may affect particularly those EQ-5D health states with tariff values (TTO-based) between 0.4 and -1.15. Hence, it seems that extreme health states of the EQ-5D system are relatively free of non-monotonic preferences, but there is a medium range of health states for which non-monotonicity may be problematic. As we showed in section 2 violations of monotonicity in duration are in conflict with multiplicative QALY models regardless their preference foundation. Moreover, even non-multiplicative QALY models as those proposed by Miyamoto (1999), *e.g.*, $H(q) \times t^{G(q)}$, that allow utility curvature to vary as a function of health state are challenged by non-monotonicity, since it needs to be monotonic in duration. Therefore, future effort should be done to try the location of MET durations across the “map” of EQ-5D health states. This would allow establishing the extent of the problem.

Our findings on preference reversals are troubling because the type of response required, the type of reasoning evoked, and even the goal perceived in the two tasks we use are similar. Both choices and rankings require a qualitative response. Both choices and rankings involve making ordinal comparisons. Finally, both choices and rankings have as primary goal to differentiate among alternatives. Thus, it is logical to expect that the two tasks evoke qualitative strategies, such as lexicographic ordering. However, preferences seem only to be lexicographic in choices but not in rankings. Given that choices are the “gold standard” for measuring preferences future research should investigate the extension of this choice-ranking discrepancy.

Although this chapter presents empirical evidence on violations of basic assumptions of QALY utility models we can also reach some positive conclusion for QALYs. First, violations of monotonicity we observe do not yield non-monotonic majority choice when data are pooled across individuals. That is, percent rate of non-monotonic preferences is always lower than 50%. Second, modal preference orderings are either increasing or decreasing in duration, never non-monotonic. Third, violations of transitivity are relatively reduced. Therefore, our findings do not imply a radical rejection to QALY models, but it qualifies positive results obtained from tests of critical assumptions of QALYs. This study suggests that violations of “structural” assumptions should be taken into account to evaluate the validity of QALY utility models.

References

- Bleichrodt, H. QALYs and HYE: Under What Conditions Are They Equivalent?. *Journal of Health Economics* 1995; 14, 17-37.
- Bleichrodt, H. and A. Gafni. Time Preference, the Discounted Utility Model and Health. *Journal of Health Economics* 1996; 15, 49-67.
- Bleichrodt, H. and M. Johannesson. The Validity of QALYs: An Experimental Test of Constant Proportional Trade-off and Utility Independence. *Medical Decision Making* 1997; 17, 21-32.
- Bleichrodt, H. and J. Quiggin. Characterizing QALYs under a General Rank-Dependent Utility Model. *Journal of Risk and Uncertainty* 1997; 15, 151-165.
- Bleichrodt, H., P. Wakker and M. Johannesson. Characterizing QALYs by Risk Neutrality. *Journal of Risk and Uncertainty* 1997; 15, 107-114.
- Bleichrodt, H. and J. L. Pinto. Loss Aversion and Scale Compatibility in Two-Attribute Trade-Offs. *Journal of Mathematical Psychology* 2002,; 46, 315-337.
- Debreu, G. *Topological methods in cardinal utility theory*, in K. J. Arrow, S. Karlin and P. Suppes (eds.) *Mathematical methods in the social sciences*. Stanford University Press 1960, 16-25.
- Fishburn, P. and P. Wakker. The invention of the independence condition for preferences. *Management Science* 1995; 41, 1130-1144.
- Keeney, R. and H. Raiffa. *Decisions with multiple objectives*, vol.1, 1976. Wiley: New York.
- Krantz, D. H., R. D. Luce, P. Suppes and A. Tversky. *Foundations of measurement*, 1971, vol 1, Academic Press: New York.
- Maas, A. and P. Wakker. Additive conjoint measurement for multiattribute utility. *Journal of Mathematical Psychology* 1994, 38; 86-101.
- Miyamoto, J. M. and S. A Eraker . A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General* 1988; 117,3-20.
- Miyamoto, J., P. P. Wakker, H. Bleichrodt and H. J. M. Peters. The Zero-condition: A Simplifying Assumption in QALY Measurement and Multiattribute Utility. *Management Science* 1998; 44, 839-849.

- Miyamoto, J. Quality-Adjusted Life Years (QALY) Utility Models under Expected Utility and Rank Dependent Utility Assumptions. *Journal of Mathematical Psychology* 1999; 43, 201-237.
- Pliskin, J. S., D. S. Shepard and M C. Weinstein. Utility functions for life years and health status. *Operations Research* 1980; 28, 206-24.
- Spencer, A. Testing the additive independence assumption in the QALY model. *New Economics Papers-HEA*, 2000.
- Sutherland, H. J., H. Llewellyn-Thomas, N. F. Boyd and J. E. Till. Attitudes towards quality of survival: The concept of maximum endurable time. *Medical Decision Making* 1982; 2, 299-309.
- Treadwell, J. R. Tests of Peferential Independence in the QALY model. *Medical Decision Making* 1998; 18, 418-428.
- Von Winterfeld, D. and Edwards, W. *Decision Analysis and Behavioral Research*, 1986. Cambridge University Press.
- Wakker, P. P. *Additive representations of preferences: A new foundation of decision analysis*, 1989. Kluwer: Dordrecht.
- Wakker, P. P. A Criticism of Healthy-years Equivalent. *Medical Decision Making* 1996; 16, 207-214.