Versión aceptada del artículo publicado en Medical Decision Making 2024, vol 44(1): 45-52
 https://doi.dox.org/10.1177/0272989X231207814

Testing non-monotonicity in health preferences

Jose-Maria Abellan-Perpiñan[1*], Jorge-Eduardo Martinez-Perez[1], Fernando-Ignacio Sanchez-Martinez[1], Jose-Luis Pinto-Prades[2]

[1] Applied Economics Department, Faculty of Economics and Business, University of Murcia, Murcia, Spain.

[2] Department of Economics, School of Economics and Business, University of Navarra, Pamplona, Spain.

*** Correspondence:**

Corresponding author: Jose-Maria Abellan-Perpiñan, dionisos@um.es

Abstract

OBJECTIVE: The main aim of this paper is to test monotonicity in life duration whatever the health status. Previous findings suggest that, for poor health states, longer durations are preferred to shorter durations up to some threshold or 'maximum endurable time' (MET), and shorter durations are preferred to longer ones after that threshold. METHODS: Monotonicity in duration is tested through two ordinal tasks: pairwise choices and rankings. A convenience sample (n=90) was recruited in a series of experimental sessions where participants had to rank-order health episodes and to choose between them, presented in pairs. Health episodes result from the combination of seven EQ-5D-3L health states and five durations. Monotonicity is tested comparing the percent rate of participants whose preferences where monotonic with the percentage of participants with non-monotonic preferences for each health state. Additionally, to test the existence of preference reversals we analyse the fraction of people who switch their preference from rankings to choices. RESULTS: Monotonicity is frequently violated across the seven EQ-5D health states. Preference patterns for individuals describe violations ranging from almost 49% with choices to around 71% with rankings. The analysis performed by separate states shows that the mean rates of violations with choices and ranking are around 22% and 34%, respectively. We also find new evidence of preference reversals and some evidence –though scarce- of transitivity violations in choices. CONCLUSIONS: Our results show that there is a medium range of health states for which preferences are non-monotonic. Extremely bad states are negative all over the duration range so you expect monotonic preferences, just as in the very good states. Our findings enlarge evidence from previous studies reporting MET preferences and introduces a new 'choice-ranking' preference reversal.

Highlights

- Our study reports significant rates of non-monotonic preferences (or 'maximum endurable time' type preferences) for different combinations of durations and EQ-5D health states. Two procedures based on ordinal comparisons are used to elicit preferences: direct choices and rankings.
- Analysis for separate health states shows that mean rates of non-monotonicity ranges from 22% (choices) to 34% (rankings), but within-subject analysis shows that non-monotonicity is even higher, ranging from 49% (choices) to 71% (rankings). These violations challenge the validity of multiplicative QALY models.
- We find that the MET phenomenon may affect particularly those EQ-5D health states that are in the middle of the severity scale, and not so much to extreme health states (i.e., very mild and very severe states).
- We find new evidence of preference reversals even using two procedures of similar (ordinal) nature. Percent rates of preference reversals range from 1.5% to 33%. We also find some (though scarce) evidence on violations of transitivity.

1 Introduction

This paper investigates, in a very basic and fundamental way, two empirical phenomena that challenge the multiplicative relationship assumed in quality-adjusted life years (QALYs) calculations, namely: non-monotonicity in life span and related preference reversals.

In the simplest case, QALYs are computed by adjusting life years (denoted as t) by the utility (v) attached to the health state (q) in which they are spent, i.e., $u(q, t) = \mathrm{v(q)} \cdot \mathrm{t}$, with u a QALY utility function over outcomes q and t, both embedded within a health episode (q, t), and v a utility function that assigns a value to every possible health state.

The correction of t by factor v(q) is simply called the QALY model[1] or the linear QALY model[2], since the utility u is linear in duration. The linear QALY model is a particular case of a more general model characterized by dropping the assumption of linearity, whereas retain that QALYs can be decomposed into a product of two separate utility functions defined over the attributes q and t , i.e., $u(\mathrm{q, t}) = \mathrm{v(q)} \cdot \mathrm{w(t)}$, where w is the function that values life duration, which can be nonlinear, and it is assumed to be increasing in duration.[3] This model is known as the multiplicative or generalized QALY model[4,5].

As apparent, the multiplicative decomposition underlying QALY calculations assumes that quantity and quality of life utilities are mutually independent. Hence, the utility of any health state is assumed to be constant, irrespective the time spent in that state. This means that for a health state valued as better than death (BTD), *i.e.*, a positive state, with $v(q) > 0$, longer durations will be preferred to shorter durations, so QALY utility $u(q, t)$ will increase monotonically with duration t, whereas if a health state is regarded as worse than death (WTD), *i.e.*, a negative state, with $v(q) < 0$, the number of

QALYs will decrease monotonically with duration. If monotonicity holds then all the QALY utility functions for life duration will have the same curvature with respect to different health states.[3]

On the contrary, if the utility of a health state depends on its duration it is no longer constant and $v(q)$ becomes $v(q, t)$, in such a way that the joint utility function $u(q, t)$ cannot be decomposed into a product of separable factors that depend, respectively, on health state and life duration, and a more general, non-multiplicative model results.[6,7]

If QALY utility does not hold a monotonic relationship with duration, the measurement of preferences for health turns more difficult. Imagine, for example, that a health state is evaluated BTD for the first 5 years, but after that the state becomes WTD. Then at least two health state utility measurements will be necessary to compute the QALYs yielded by health episodes longer than 5 years: one before the 'switching' time point beyond which the respondent no longer prefer to live more years to fewer, and another one after that point.[8] This example illustrates indeed the phenomenon coined as "maximum endurable time" (MET)[9], which is a particular case of non-monotonic preferences. Put in graphical terms, it can be depicted by an inversed-U-shaped QALY utility function with a single peak at a time point, *i.e.*, the MET, beyond which (5 years in our example) the health state is seen as increasingly intolerable.[10]

Nevertheless, and despite that the findings reported by many studies[10,11,12,13,14] have been commonly interpreted as supporting the hypothesis of the existence of the MET, there are still various issues that require to be elucidated. Firstly, as noted before, MET preferences are just one example of non-monotonic preferences. Think of, instead of the typical curved pattern of MET preferences, with first upward and next downward sloping sections, just the opposite pattern: that described by a U-shaped curve. This non-monotonic pattern was predominant among the respondents that were found to violate monotonicity in a study.[15] Around 30% of the sample valued WTD increasingly over time, which is contradictory with the conventional MET. So, it is worthy to deep inside the “map” of diverse non-monotonic preferences, which is one of the aims of this paper.

Secondly, the disparity between results such as those we have just described above and those found by the majority of the remaining studies “may be due to differences in the way in which MET is assessed across studies”[15] (p. 400). Most studies[10,11,12,13] have tested MET preferences by means of the comparison of just one direct choice between health episodes of type (q, t_1) and (q, t_2), with $t_1 < t_2$, and the implied choice derived from time trade-off (TTO) assessments for the same episodes. A preference reversal typically arises from this comparison: respondents prefer the episode with the shorter duration when asked directly, but assign with the TTO more utility to the episode with the longer duration.

Faced with this disparity, researchers[13] concluded that the preference reversal “hides the MET preferences when values are assessed with the time trade-off task” (p. 495). The explanation given to this preference reversal is attributed to a “rule of thumb” followed by respondents when answering TTO questions, called the proportional heuristic.[10,11] Shortly, this heuristic means that respondents choose a duration in full health as a fixed proportion of the duration in the poor health state. Therefore, health state utility remains roughly constant irrespective the duration used as stimulus in TTO measurements, seemingly confirming the QALY model, though respondents’ preferences are not actually time-independent.[12] According to several authors[10,11,12], the use of this heuristic is driven by scale compatibility. This compatibility effect states that respondents weigh more heavily the stimulus attribute more compatible with the response scale[16], and it is one of the explanations to the so-called ‘choice-matching’ discrepancy.[17,18] As in the TTO individuals provide life years as a response, then life duration will receive a larger weight than that for the health state, which could lead to neglect that, because of the poor health state, fewer years should be preferred to more. This fact has led to claim[10] that, at least for severe health states, the usage of the TTO is not appropriate.

For all the reasons mentioned, this paper aims three objectives: (1) to test unambiguously non-monotonic preferences by means of a variety of direct choices encompassing an ample set of different health episodes, including death. Since various health episodes are used, we also analyse possible intransitive preference orderings by inspection of the series of direct choices made by respondents. Furthermore, participants in the study also rank the same health episodes, which provides a parallel way to check non-monotonic patterns; (2) to verify if non-monotonic patterns are a function of severity and/or the type of task used; and (3) to test whether preference reversals, in the presence of non-monotonic patterns, may arise even if no matching task is used. The use of choices and rankings allow us to test potential preference reversals across both tasks. Note that the response scale of the two procedures is similar, i.e., choose one episode over another or rank an array of them, so scale compatibility should not provoke a discrepancy between both. .

The paper is structured as follows. Section 2 describes the experiment conducted to test failures in monotonicity and potential preference reversals between direct choices and rankings of the same set of chronic health outcomes. Results are provided in section 3. A discussion closes the paper.

2 The experiment

2.1 Participants and experimental sessions

Participants were 90 Economics undergraduate students who participated for course credits. They were recruited by means of a participation call posted in the teaching digital platform of the University of Murcia. No additional incentives were provided, apart for the course credits.

Each participant attended three experimental sessions, one to rank-order chronic health episodes (ranking session) and the other two to choose between them (choice sessions). The tasks asked in each session were administered by paper-based booklets. The sessions were run by one of the authors in small groups with at most five subjects at a time in a behavioral laboratory at the University of

Murcia. In order to avoid order and memory effects, tasks within sessions were randomly assigned to participants, and sessions were separated by one week each. Each session lasted at most 40 minutes.

2.2 Chronic health episodes

We used seven health states based on the EQ-5D-3L classification system[19]. According to this system health states are described by means of five dimensions, each of which can take one level out of three possible. Table 1 shows the description of the health states, anonymously labelled T-Z.

The health states were chosen to cover the range of the value set generated by the EQ-5D-3L algorithm for Spain[20]. According to this algorithm, the values attached to each of the health states are 0.91, 0.54, 0.43, 0.25, -0.14, -0.44, and -0.65, for states T(11112), U(11113), V(11312), W(12223), X(13332), Y(33232), and Z(33333), respectively. Our selection encompasses one 'very mild' state (11112), two 'mild' states (11113 and 11312), one 'moderate' state (12223), two 'severe' states (13332 and 33232), and the worst possible state that the EQ-5D-3L system can describe (the 'pits' state 33333) [21].

From the combination of each health state with durations 0, 13, 24, 38, and 57 years respectively, we obtained the five health episodes per state presented to participants. Previous studies investigating MET preferences that have used EQ-5D-3L health states, included in their designs durations up to a maximum of 20 years.[14] As Scalone et al.[22] argue it is interesting, therefore, to explore how duration affects the preference for health states beyond that time horizon. For this reason, we included longer durations with a maximum duration of 57 years, so as not to exceed the life-expectancy of participants (mean age was twenty years). In addition, we intentionally avoided using "round" durations, e.g., 10, 20, 30 years, in an attempt to enhance respondents' deliberation to compare the different episodes.

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

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

2.3 Tasks

Prior to the first experimental session, subjects were introduced to the EQ-5D system. In addition, at the beginning of each session the participants made choices and rankings that could mean preferring less to more years in the same health state. The questionnaires began with a trial question that was checked with participants before starting the experiment.

Seven rankings (one per health state) of five possible durations were obtained from each participant. So, for example, for state T episodes (T, 0 years), (T, 13), (T, 24), (T, 38) and (T, 57) are ranked. Episodes were printed on a set of cards which, to avoid order effects, were distributed at random. Each episode was described by means of a short sentence, e.g., ‘You are living 38 more years in health state T’. To avoid response errors, participants were asked to confirm their rankings. If they did not confirm it, they could change the ordering. We repeated the process until participants did agree with the orderings revealed. After that, participants were asked to fill in a table, where they had to write, for each health state, the position 1 to 5 that corresponded to each duration, from most to least preferred episode.

In the choice sessions, participants were asked to make choices between two chronic health episodes. As there are five different durations, ten pairs of health episodes for each EQ-5D health state follow. Overall, each participant made seventy choices, i.e., 10 pairs × 7 health states, evenly distributed across the two questionnaires administered in each session. The order in which choices were presented within each questionnaire was random. To avoid response errors, participants were asked to confirm their choices by filling in a table, where they have to write down their choice for every pairwise comparison. The table was made of four columns, the first two showing the two options for each pairwise comparison, under the headings ‘Alternative 1’ and ‘Alternative 2’ (e.g. 24 years in health state U vs. 38 years in health state U). The other two columns offered two possibilities to participants: ‘I choose Alternative 1’ and ‘I choose Alternative 2’. Respondents had to tick the chosen option. This additional task forced them to check earlier responses.

2.4 Analyses

As noted in Introduction, multiplicative QALY models imply that preferences should satisfy monotonicity in duration, which means that for all (q_1, t_1), (q_1, t_2) with $t_2 > t_1$ either (q_1, t_2) is ‘strictly preferred to’ (henceforth denoted by the individual strict preference relation $\succ$) (q_1, t_1), i.e., increasing monotonicity, or $(q_1, t_1) \succ (q_1, t_2)$, i.e., decreasing monotonicity. Likewise, it is also assumed that preferences satisfy transitivity, i.e., if $(q_1, t_1) \succcurlyeq (q_1, t_2)$ and $(q_1, t_2) \succcurlyeq (q_1, t_3)$, then $(q_1, t_1) \succcurlyeq (q_1, t_3)$, with $\succcurlyeq$ denoting the weak preference relation ‘at least as preferred as’. Since rankings force pairwise comparisons to be consistent while simple choices do not, violations of transitivity were analysed in the choice task only by inspection of intransitive cycles, *i.e.*, $(q_1, t_1) \succcurlyeq (q_1, t_2)$ and $(q_1, t_2) \succcurlyeq (q_1, t_3)$ but also $(q_1, t_3) \succcurlyeq (q_1, t_1)$. All the analyses were performed by health state.

To achieve the first objective of this paper, incidence of non-monotonic and intransitive preferences was analysed in two ways. On the one hand, participants' responses were classified into one of the different preference patterns observed in the data. That is, we counted the number of participants with non-monotonic or intransitive preferences for each health state q_i and procedure, i.e., choices and rankings. Participants whose preferences were non-monotonic for at least one health state (e.g. a respondent with monotonic preferences for, say, four states, and non-monotonic for the remaining three states) were classified as non-monotonic subjects. MET patterns and opposite non-monotonic patterns, i.e., those revealing that shorter durations in a given health state are ranked as WTD and longer durations as BTD, were differentiated where applicable as non-monotonic MET preference patterns and other ones. On the contrary, those respondents who exhibited monotonic preferences for all the states were classified as exclusively increasing, exclusively decreasing or both increasing and decreasing monotonic ones, depending on they always preferred more to fewer years, fewer to more years, or they preferred more to fewer years for some states, and the opposite pattern for other ones. Subjects with intransitive cycles for one or more states were classified as intransitive ones.

On the other hand, we also calculated both the percent rate P(m) of participants for whom preferences were monotonic and the percent rate P(non-m) of participants with non-monotonic preferences, for each health state q_i and task. The magnitude of P(non-m) in regards to P(m) gives, in this way, an idea of its relative frequency. The same was done to inspect intransitive cycles in the choice task: percent rate P(t) of participants for whom preferences were transitive and percent rate P(i) of participants with intransitive preferences are calculated for each health state as well.

To verify if monotonicity is the most frequent pattern, i.e., the ‘modal’ one, we tested, for each health state q_i and task, whether P(m) > P(non-m) holds. Those participants who exhibited intransitive preferences in the choice task for any of the health states, were excluded from the test of monotonicity. Monotonocity was tested by using the goodness-of-fit Chi-squared (χ^2) test.

To fulfil the second objective, i.e. whether non-monotonic patterns change depending on the severity and/or the type of task used, 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.

Lastly, the existence of preference reversals (third aim of the paper) was analysed by calculating the percent rate of preference reversals for each health state as the fraction of people who switch their preference from rankings to choices. That is, respondents who, in a direct choice, preferred the health state with duration t_i over the same outcome with a duration t_j, but ranked a t_j duration above a t_i duration in the rank-ordering task for the same health state. The rates were computed both with and without participants who yielded any intransitivity.

3 Results

In regards to the first aim of the paper, i.e., to test non-monotonicity in duration, only six participants in the choice task and one participant in the ranking task display increasing monotonic preferences for all health states. The pattern is "mixed" (i.e. increasing monotonic preferences for some health states and decreasing monotonic preferences for others) for 20 participants in the choice task and 22 in the ranking. It is also found that most participants display non-monotonic MET preferences for at least one health state. As can be seen in Table 2, forty-three participants (47.8% of participants) behave according this pattern in the choice task and sixty-four (71.1%) in the ranking task. There were only four participants (one in the choice task and three in the ranking task) describing, for some health state, a non-monotonic pattern contradictory with MET predictions, reported in Table 2 as the category 'Other'. Subjects included in this category are dropped in subsequent analyses.

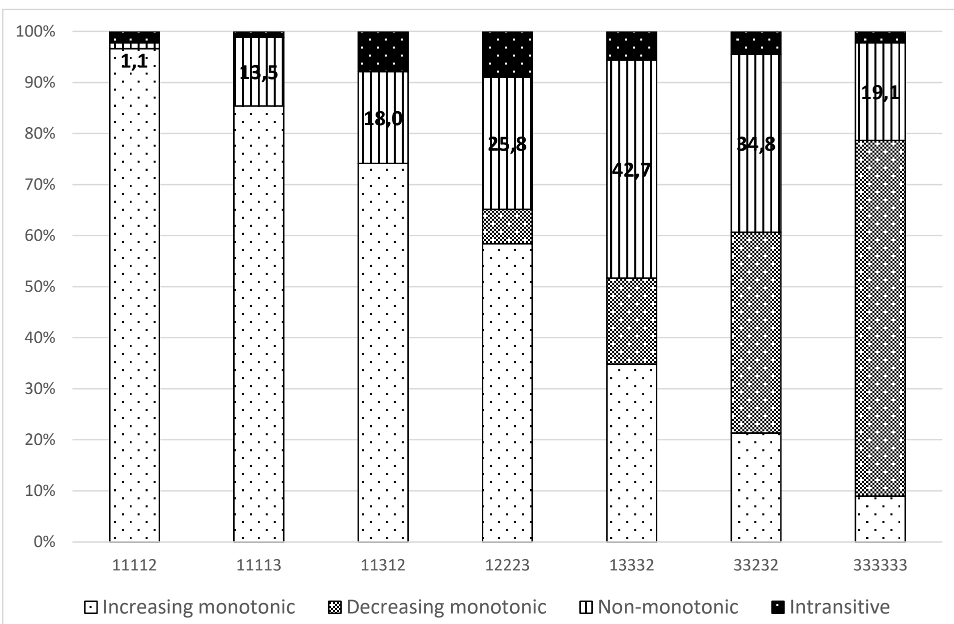
Table 2. Preference patterns for individuals in choice and ranking

Preference pattern	Choice	Ranking
Exclusively increasing monotonic	6	1
Exclusively decreasing monotonic	--	--
Both increasing and decreasing monotonic[1]	20	22
Non-monotonic MET[2]	43	64
Intransitive[3]	20	--
Other[4]	1	3

Note: [1]: preferences are increasing monotonic for some health states and decreasing monotonic for others; [2]: preferences are non-monotonic for at least one health state (only one participant displayed non-monotonic preferences for all health states) according to the 'maximum endurable time' (MET) pattern; [3]: preferences are intransitive for one or more health state; [4]: preferences are non-monotonic but not follow the MET pattern

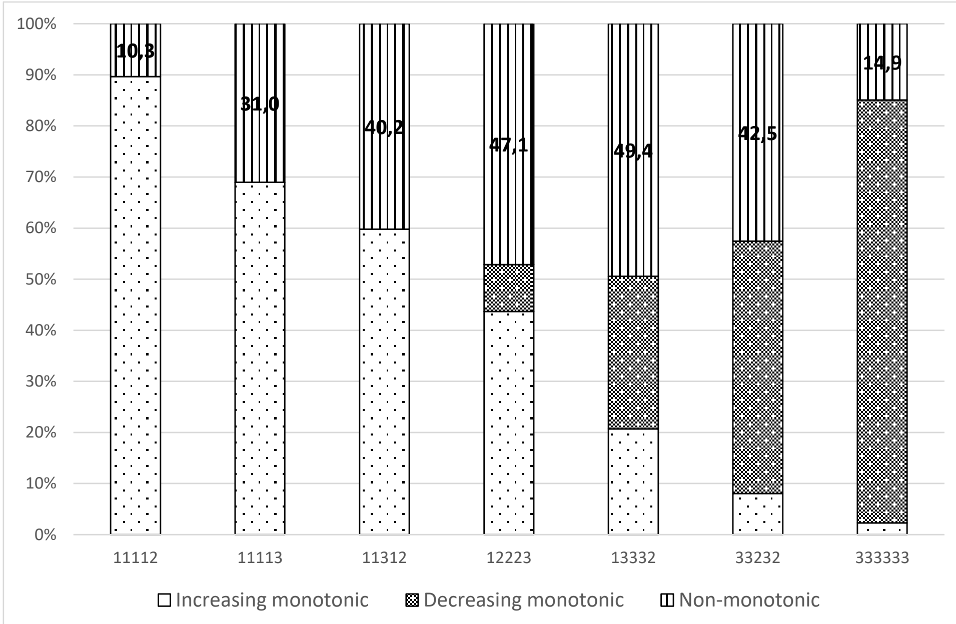
Twenty participants made intransitive choices in the choice task at some point. After removing these participants, the percent rate of non-monotonic MET preferences is very similar in the ranking task (66.7%) and somewhat higher in choices (62.3%).

Figure 1. Preference patterns for separate health states, choice task (percent rate) (N = 89)*



*: one participant who showed non-monotonic, non-MET, preferences for state 12223, is omitted from the calculus.

Figure 2. Preference patterns for separate health states, ranking task (percent rate) (N = 87)*



*: three participants who showed non-monotonic, non-MET, preferences for states 11112 and 12223, are omitted from the calculus.

Four main points arise from the inspection of figures 1 and 2. First, percent rates of non-monotonic MET preferences range from 1.1% (state 11112) to 42.7% (state 13332) under the choice task, and from 10.3% to 49.4% (for the same states) under the ranking task. Second, percent rate of non-monotonic MET preferences increases with severity, reaching its maximum for health state 13332. Third, we observe that percentages of non-monotonic MET preferences are lower for choices than for rankings. Fourth, percent rates of intransitivities are relatively small. They range from 1.1% for health state 11113 to 9% for health state 12223.

It can be seen that as the severity of health states increases the number of subjects who prefer longer over shorter durations decreases. In the case of very severe health states (33232 and 33333) preferences are negatively monotonic since shorter duration are preferred to longer ones.

After excluding participants with intransitive responses[1], we observe that, under the choice task, the rate of monotonic preferences is significantly higher than the rate of non-monotonic MET preferences in all cases except for health state 13332 (Chi-square, $p = 0.093$), so non-monotonicity is, for that state, almost as likely as monotonicity (39.1% vs. 60.9%). Furthermore, although for the remaining states discrepancies between monotonic and non-monotonic MET 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.*, 27.5% and 34.8%.

Results from the ranking task show more robust evidence contrary to monotonicity in duration. In particular, we do not find significant differences between monotonic and non-monotonic MET rates for health states 12223, 13332, and 33232 (Chi-square, $p = 0.337$, $p = 0.471$, and $p = 0.092$, respectively). Percent rates of non-monotonic MET preferences for these states are 43.5%, 44.9%, and 39.1% respectively. They are also high for health states 11113 (29%) and 11312 (34.8%), although monotonicity cannot be rejected.

With respect to the second objective of this paper, i.e., to verify if non-monotonic patterns are a function of severity and/or task, it is apparent in figures 1 and 2 that monotonicity is more frequently violated with rankings than with choices. Indeed, we find that the probability of exhibiting non-monotonic MET 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 occurrence of non-monotonic MET preferences is not independent on the health status (Cochran Q test, $p < 0.0001$ for both ranking and choice tasks). The percent rate of non-monotonic MET preferences increases with severity level from health state 11112 to state 13332, for which the highest rate is reached. Moreover, the inspection of individual responses

[1] The results hardly change when we keep the 20 subjects with intransitive preferences.

suggests that the most preferred duration by the participants is shorter as the severity increases. In other words, the MET moves to the left (i.e. shorter durations) as severity increases. Lastly, regarding our third objective, i.e., to test preference reversals across tasks, the proportion of preference reversals between the rank ordering and choice tasks was 1.5%, 19%, 24.9%, 33%, 22%, 13.5%, and 6.2% for health states 11112, 11113, 11312, 12223, 13332, 33232, and 33333 respectively. On average, intransitivities explain less than 5% of these reversals.

4 Discussion

4.1 Main findings

We use two different procedures to elicit preferences: choices and rankings. We find that monotonicity is frequently violated. Preference patterns for individuals reveal that violations of monotonicity range from around 48% with choices to 71% with rankings. Analysis for separate health states shows that the rate of violations for some health states is near 50% in the ranking task. We observe that violations of monotonicity increase with severity, and are higher for the states 12223 and 13332 than for more severe states, such as 33232 and 33333.

We find new evidence of preference reversals with two choice-based procedures. Percent rates of preference reversals range from 1.5% for health state 11112 to 33% for state 12223. Finally, we also find some (though scarce) evidence on violations of transitivity.

4.2 Previous related studies

MET preferences were first reported by Sutherland et al[9]. These researchers found that in highly dysfunctional health states, the proportion of respondents preferring death increased as duration of survival in those states got longer. However, only comparisons with immediate death cannot identify by themselves the ‘switching’ time point beyond which shorter durations are preferred to longer durations. This also requires to include choices between identical health states of different duration, as in the experiment presented in this paper.

Dolan[23] estimated EQ-5D tariff based on VAS valuations for 42 EQ-5D states and three different durations. Utility estimate for a given health state is a decreasing function of both its severity and its duration, in such a way that even for milder states, utility decreases with duration. This finding contrasts with recent estimations of QALY utilities for different health episodes[22,24] that show that utility declines with duration for severe problems, but not for milder and extreme problems, for which utility increases (or disutility decreases) but a decreasing pace. Our results are in line with these studies, suggesting that extremely bad states are negative all over the duration range, just as very good states are positive EQ-5D states, whereas there is a medium range of health states, i.e., moderate and severe ones, throughout preferences are frequently non-monotonic.

We find percent rates of non-monotonicity for health state 13332 close to that reported by Dolan and Stalmeier[12] for EQ-5D state 21223, the single state they consider. On the contrary, our results suggest that rates of non-monotonic preferences for health states 12223, 13332, 33232, and 33333 are higher than those reported by other studies[10,11] that have used only one direct choice and two TTO questions to test monotonicity in preferences. All these authors report preference reversal rates significantly higher (ranging from 74% to 86%) than those we find across choices and ranking comparisons. Hence, it seems that the use of two tasks with a similar response scale may make preference reversals less substantial, though remains important and systematic. This finding is a novelty in the domain of health outcomes, using health episodes entirely riskless, that adds to previous evidence reported by studies also using choice-based procedures, but applied to risky health outcomes.[25,26]

Robinson and Spencer[15] reported a majority of violations of monotonicity with patterns opposite to that predicted by MET. This evidence comes from the observation of utility estimates for different combinations of durations with EQ-5D health state 23323. Utilities for health episodes were elicited by applying a modified TTO procedure, initially called a ‘life profile’ approach, and that later on will be known as a lead TTO.[27] As described before, the presence of non-monotonic patterns distinct to those consistent with MET preferences are scarce in our data. The only four violations of monotonicity reported in this paper in a direction contrary to that predicted by MET seem to be respondents' mistakes rather than true preferences. Therefore, the MET hypothesis is consistently supported by the data analysed here, with the added value that it has been checked via simple preference questions, without using any variant of the TTO. Moreover, evidence reported in this paper encompasses a wide severity range, including seven different EQ-5D states, and not only one, as Robinson and Spencer[15] used.

The study conducted by Stalmeier et al.[14] is, to the best of our knowledge, that closest to ours. The authors used two series of direct choices to test MET preferences. On the one hand, choices between a health state of a specified duration and death, and, on the other hand, choices between two identical states of different duration. Proportions of individuals with preferences consistent with MET predictions were similar with both types of choices, occurring more frequently for severe health states. The percent rates of non-monotonic preferences reported in their paper do not exceed 30% for any of the five EQ-5D states they consider, whereas we find rates higher rates for some states. Nevertheless, the qualitative picture is similar in the two studies, though non-monotonic preferences are more frequent in our data. Note that experiment protocols, nature of the sample and set of health states are different in both studies.

As Miyamoto et al.[3] assert, the phenomenon of MET for a given health state constitutes a basic counter-example to the multiplicative QALY model. Our data clearly show that the time point of the MET moves to the left as severity increases, indicating therefore that QALY utility functions for life durations have a different curvature with respect to different health states, something that contradicts mutual utility independence between life duration and quality of life. A complementary result is reported by Attema and Brouwer[8], who found stronger discounting of WTD states than BTD states, which also contradicts the multiplicative QALY model.

Preference reversals observed in this paper are particularly troubling, because they cannot be explained by compatibility effects, such as those concerning the usual 'choice-matching' discrepancy reported between direct choices and TTO responses.[10] So a 'choice-ranking' discrepancy arises from our data, similar to that previously identified by Bleichrodt and Pinto[25] for risky treatments. The different domain of the health outcomes used in their study (risky) and ours (riskless) makes that explanation to preference reversals hypothesized by these authors (i.e., anticipation of disappointment and elation in risky choice) is not valid for our data.

Although intransitive preference ordering has been suggested as an explanation to the classical 'choice-matching' discrepancy[28], later evidence suggest that intransitivity is likely to explain only 10-20% of the phenomenon[29]. Our data supports this observation also for preference reversals between choice and ranking, since intransitivity hardly explain 5% of them.

A possible explanation for our findings can be the so-called evaluability hypothesis[30]. According to this hypothesis, the way in that attributes are evaluated, separately or jointly, provides a different information to subjects that can be led to preference reversals. In our experiment, durations for each health state are compared together (joint evaluation) in rankings while they are compared head to head (something closer to a separate evaluation) in pairwise choices, so a preference reversal might arise between these two different "evaluation" modes. Joint evaluation of health episodes can make respondents more conscious of the interaction between duration and health state, whereas separate evaluation can obscure that relationship, making duration more salient. In this way, non-monotonicity would be more frequent in ranking than in choice, like indeed our results reveal.

4.3 Limitations

This study is not exempted from limitations. First, assuming that, in general, students are in good health, their perception of the severity of a hypothetical poor health state may differ from that of older (i.e. less healthy) people because they never experienced adaptation to a health problem. Other objection may concern the sample-size used, although it is larger than others used in some previous studies[11,12,13,31]. Participants in our experiment did not receive financial compensation. Instead, participation in the experimental sessions were rewarded with course credit. Though it would be interesting to check if results are robust to changes in compensation, we do not believe that financial motivation may vary our findings[32]. On another note, 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. Another objection could be that the health episodes used were too simple, inducing easily salience-based decision. However, if this had been the case, we believe that there would not have been so many violations of monotonicity as we observe. Lastly, it could be argued that participants in our experiment might have been found it hard to perceive living for very long durations. For this reason, analyses were carried out after leaving out 57 years duration. Rates of non-monotonicity decrease for all health states, although non-monotonic preferences persist systematically.

4.4 Implications

From our study it can be inferred that the MET phenomenon may affect particularly those EQ-5D health states that are in the middle of the severity scale. Therefore, it may be necessary to explore the role of non-multiplicative models to describe non-monotonic interactions between duration and health quality.

Our findings on preference reversals are troubling because choices and rankings have many similar features[18,33]. Thus, it is logical to expect that there is no difference in information processing strategy between the two tasks, and that compatibility effects do not cause preference reversals. However, in our data, non-monotonic preferences seem to be more likely in rankings than in choices. We hypothesize that this choice-ranking discrepancy may be due to information effects driven by the different evaluation mode (joint vs. separate) induced in each task. So, future research should test this hypothesis by, for example, comparing a choice-based ranking task[25], according to which respondents are asked to choose the most preferred health episode, next the second one, and so on, to a conventional ranking. In addition, it would be interesting to confront respondents to their choices and rankings and ask them the reasons why have performed such preference orderings.

Although this paper presents evidence on monotonicity violations we can also reach some positive conclusion for QALYs. First, when data are pooled across individuals, the percent rate of non-monotonic preferences is not higher than 50%. Second, at the aggregate level, modal preference orderings are either increasing or decreasing in duration, never non-monotonic. Third, violations of transitivity are relatively small (between 1.1% and 8.9%, depending on the health state). Therefore, our findings do not imply a radical rejection to the QALY models, but they qualify its descriptive ability.

Acknowledgments

Financial support for this study was provided entirely by a grant from the Spanish Ministry of Economy, Industry and Competitiveness, grant PID2019-104907GB-I00, and a grant from the Seneca Foundation (Science and Technology Agency of the Region of Murcia), grant 20825/PI/18. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The authors acknowledge administrative support given by administrative staff of the Faculty of Economics and Business of the University of Murcia (Spain)

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1]Bleichrodt H, Wakker P, Johannesson M. Characterizing QALYs by Risk Neutrality. J Risk Uncertain. 1997;15:107–14.

[2]Abellán-Perpiñán JM, Pinto-Prades JL, Méndez-Martínez I, Badía-Llach X. Towards a better QALY model. Health Econ. 2006;15(7):665–76.

[3]Miyamoto JM, Wakker PP, Bleichrodt H, Peters HJM. The Zero-Condition: A Simplifying Assumption in QALY Measurement and Multiattribute Utility. Management Science. 1998; 44(6):839-849.

[4]Miyamoto JM, Eraker SA. A multiplicative model of the utility of survival duration and health quality. undefined. J Experimental Psycohology: General. 1988;117(1):3–20.

[5]Attema AE, Brouwer WBF. The value of correcting values: influence and importance of correcting TTO scores for time preference. Value Health. 2010;13(8):879–84.

[5]Attema AE, Brouwer WBF. The way that you do it? An elaborate test of procedural invariance of TTO, using a choice-based design. Eur J Heal Econ. 2012;13(4):491–500.

[6]Miyamoto JM. Quality-Adjusted Life Years (QALY) Utility Models under Expected Utility and Rank Dependent Utility Assumptions. Journal of Mathematical Psycology. 1999;43:201-237.

[7]Attema AE, Bleichrodt H, Wakker PP. A direct method for measuring discounting and QALYs more easily and reliably. Med Decis Mak. 2012;32:588-593.

[8]Attema AE, Brouwer WBF. A test of independence of discounting from quality of life. J Health Econ. 2012;31:22-34.

[9]Sutherland HJ, Llewellyn-Thomas H, Boyd NF, Till JE. Attitudes Toward Quality of Survival: The Concept of "Maximal Endurable Time." Med Decis Mak. 1982;2(3):299–309.

[10]Stalmeier PFM, Wakker PP, Bezembinder TGG. Preference reversals: Violations of unidimensional procedure invariance. G. Balint, Antala B, Carty C, Mabieme J-MA, Amar IB, Kaplanova A, editors. J Exp Psychol Hum Percept Perform. 1997;23,:1196–205.

[11]Stalmeier PFM, Bezembinder TGG, Unic IJ. Proportional heuristics in time tradeoff and conjoint measurement. Med Decis Mak. 1996;16(1):36–44.

[12]Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. J Health Econ. 2003;22(3):445-58.

[13]Stalmeier PFM, Chapman GB, De Boer AGEM, Van Lanschot JJB. A fallacy of the multiplicative

qaly model for low-quality weights in students and patients judging hypothetical health states. Int J Technol Assess Health Care. 2001;17(4):488–96.

[14]Stalmeier PFM, Lamers LM, Busschbach JJV, Krabbe PFM. On the assessment of preferences for health and duration: Maximal endurable time and better than dead preferences. Med Care. 2007;45(9):835–41.

[15]Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ. 2006;15: 393-402.

[16] Slovic P, Griffin D, Tversky A. 1990. Compatibility effects in judgment and choice. R. Hogarth, Ed. Insights in Decision Making: A Tribute to Hillel J. Einhorn. University of Chicago Press, Chicago, IL. 5–27.

[17] Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. Psychological Review. 1988;95:371‑384.

[18]Fischer GW, Hawkins SA. Strategy Compatibility, Scale Compatibility, and the Prominence Effect. J Exp Psychol Hum Percept Perform. 1993;19(3):580–97.

[19]Brooks R, De Charro F. EuroQol: the current state of play. Health Policy (New York). 1996 Jul 1;37(1):53–72.

[20]Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Mak. 2001;21(1):7–16.

[21] Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. 1995; Discussion Paper 138. York: Centre for Health Economics.

[22]Scalone L, Stalmeier PFM, Milani S, Krabbe PFM. Values for health states with different life durations. Eur J Heal Econ. 2015;16(9):917–25.

[23]Dolan P. Modelling valuations for health states: the effect of duration. Health Policy (New York). 1996 Dec 1;38(3):189–203.

[24]Craig BM, Rand K, Bailey H, Stalmeier PFM. Quality-Adjusted Life-Years without Constant Proportionality. Value Heal. 2018;21:1124–31.

[25]Bleichrodt H, Pinto Prades JL. New evidence of preference reversals in health utility measurement. Health Econ. 2009;18(6):713–26.

[26]Oliver A. Further evidence of preference reversals: Choice, valuation and ranking over distributions of life expectancy. J Health Econ. 2006;25(5):803–20.

[27]Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the ‘‘lead time’’ approach. Health Econ. 2011; 20(3):348–61.

[28]Loomes G, Sugden R. A rationale for preference reversal. American Economic Review. 1983; 73: 428‑32.

[29]Tversky A, Slovic P, Kahneman D. The causes of preference reversals. American Economic Review.

1990;80 (1):204–217.

[30]Hsee CK, Blount S, Loewenstein GF, Bazerman MH. Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. Psychol Bull. 1999;125(5):576–90.

[31]Gudex C, Dolan P. Valuing health states: the effect of duration. 1995; Available from: https://www.york.ac.uk/media/che/documents/papers/discussionpapers/CHE Discussion Paper 143.pdf

[32]Mellers BA, Ordóñez LD, Birnbaum MH. A change-of-process theory for contextual effects and preference reversals in risky decision making. Organ Behav Hum Decis Process. 1992;52(3):331–69.

[33]Pinto J, Sanchez F, Abellan J, Martinez J. Reducing preference reversals: The role of preference imprecision and nontransparent methods. Heal Econ. 2018;27(8):1230–46.