Making composite time trade-off sensitive for worse-than-dead health states

Michał Jakubczyk[1], Bram Roudijk[2], Stefan A. Lipman[3], and Peep Stalmeier[4]

[1] *SGH Warsaw School of Economics, Poland; ORCID=0000-0002-0006-6769*

[2] *EuroQol Research Foundation, Rotterdam, the Netherlands*

[3] *Erasmus School of Health Policy & Management, Erasmus Centre for Health Economics Research, Erasmus University Rotterdam, the Netherlands*

[4] *Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands*

July 10, 2024

Abstract

Objective The utilities elicited with the composite time trade-off (cTTO) method for health states worse-than-dead (WTD) often correlate poorly with other severity measures, indicating a poor sensitivity of cTTO. We aimed to explore modifications to cTTO to better understand this phenomenon and identify potential improvements.

Methods 480 respondents completed an online TTO interview, each valuing 12 EQ-5D-5L health states. The participants were divided into four arms, A–D. Arm A followed the standard cTTO, serving as a reference. In arm B, we removed the sorting question comparing immediate death vs. 10 years in a valued state. Arm C allowed for utility values < -1 by reducing the time in the valued state in lead-time TTO (LT-TTO) part of cTTO. In arm D, we randomly choose the starting negative utility in LT-TTO. Utility value distributions, correlations between utilities and level sum score (LSS), and inconsistencies between Pareto-ordered states were analysed.

Results Arm A replicated the lack of significant correlation between LSS and the negative utility observed in previous work. Of the experimental arms, only arm B exhibited a significant negative correlation. Compared to arm A, arm B produced a higher proportion of WTD states (46.5% vs. 26.3%), less negative utility for WTD states on average (-0.571 vs. -0.752), and a lower mean censored utility for 55555 (-0.486 vs. -0.406).

Conclusion The observed lack of correlation between LSS and utility for WTD states appears linked to the use of comparison with immediate death. LT-TTO is capable of eliciting utility values in a way that is sensitive to severity. Modifying the initial questions in cTTO to identify if health states are BTD or WTD could be considered.

Keywords: health states; worse than dead; EQ-5D-5L; sensitivity; composite time trade-off

1 Introduction

To measure the health benefits of health technologies, a QALY model is often used (QALY stands for quality-adjusted life years). In this model, the health states are assigned values, *index weights*, which are subsequently multiplied by the number of years spent in a given health state. In cost-utility analysis of health technologies (CUA), health states are usually defined using one of the EQ-5D family of instruments [Kennedy-Martin et al, 2020]. In these instruments, a state of health is defined using five dimensions: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). Each dimension is assigned a level (either 1–3 for EQ-5D-3L or 1–5 for EQ-5D-5L), which represents the amount of health problems, where 1 denotes no problems and the last level (either 3 or 5) represents extreme problems. Henceforth, we focus on EQ-5D-5L, which was used in the present study.

The index weights are derived in valuation studies in which the preferences for health states of a given society are elicited [for instance, see Versteegh et al, 2016; Golicki et al, 2019; Pickard et al, 2019]. The preferences of each individual are expressed using von Neumann and Morgenstern utilities (1944) with additional assumptions that allow the multiplicative form in the QALY model [Bleichrodt et al, 1997; Miyamoto et al, 1998]. The utilities are scaled in a way to make the utility of being dead equal to 0, and the utility of full health equal to 1, referred to as *QALY scale*. Health states that are considered worse than dead (WTD) receive negative utility in the QALY model.

The elicitation of preferences for EQ-5D health states typically uses the EQ-VT protocol [Stolk et al, 2019]. It comprises two elicitation tasks: composite time trade-off (composite TTO, cTTO) and discrete choice experiment (DCE). In cTTO, the respondent compares a shorter life in full health with a longer life that includes health problems: the respondent trades-off the life years in full health until the two lives seem equally attractive. In DCE (as implemented in EQ-VT), health states are compared without duration being specified or immediate death being used as one of the alternatives. Such DCE tasks alone cannot produce utility values on the QALY scale, and, as such, cTTO is crucial for the anchoring of utilities, i.e., the positioning of index weights on the QALY scale.

Yet, there are doubts as to whether cTTO is sensitive enough to capture severity for health states considered WTD. In other words, there are doubts whether for such states the utility values produced by cTTO change meaningfully for states that appear to be more or less severe. The discussion originated from an observation made in Gandhi et al [2019], who showed that negative utility elicited with cTTO is very poorly correlated with level sum score (LSS), i.e., a crude, non-preference-based, measure of severity that uses a simple sum of levels for all five dimensions. In EQ-5D-5L, the LSS values can take a value between 5 (for health state 11111, which represents no problems in any dimension, hence, full health) and 25 (for health state 55555).

Subsequently, Roudijk et al [2022] pointed out that the correlation (between LSS and negative utility, omitted henceforth for brevity) may be shrunk towards zero in view of analysing a subset of the data based on the utility value [see Hausman and Wise, 1977]. Roudijk et al [2022] suggested the data should be analysed separately for the subgroups of respondents defined using how many states they considered WTD. However, Jakubczyk [2023] showed that the results of the split analysis do not change even after the negative utilities are reshuffled to guarantee insensitivity: hence, the conclusion of sensitivity based on the split analysis is invalid. In that study, it was also shown that while the use of the subset does shrink the correlation, a non-zero correlation should still be present if cTTO was sufficiently sensitive.

In the literature, various properties of cTTO were discussed that may explain the lack of correlation. Jakubczyk et al [2023] suggested that it may be the result of how the cTTO is composed of two slightly different tasks (see Section 2.2 for details): a regular TTO for states that are better than dead (BTD) and a lead-time TTO (LT-TTO) for WTD states. Which of the two is used depends on respondents' preferences in a question embedded in the start of every cTTO task: comparing living in the to be valued health state for 10 years with the alternative immediate death. If the former is preferred, a standard TTO task is conducted (comparing 10 years in impaired health to some years in full health), whereas if the latter is preferred, then a state is considered WTD and LT-TTO starts. In LT-TTO, life years in full health are added to both alternatives to make further trade-offs possible. Seeing as the comparison between 10 years in impaired health vs. immediate death sorts the respondents into groups undertaking a task for health states BTD and WTD, we refer to this comparison as the *sorting question*. For some respondents, it may be appalling to choose immediate death in the sorting question, while it may be acceptable to trade off years in LT-TTO. In consequence, only severe states will be considered WTD and subject to LT-TTO, respondents will avoid living in these states in LT-TTO by trading off life years, and few just slightly negative utilities will be observed which may contract the range of negative values and reduce the sensitivity.

Earlier work has shown that cTTO suffers from strong left-censoring, with about 10% observations ending up at the left end of the available range [Liao et al, 2023]. Such censoring may explain the poor sensitivity: when many different health states all end up receiving utilities of -1 because of the task construction, states that are in fact considered by respondents as worse do not receive a lower utility, which reduces the sensitivity.

Another feature of cTTO that may explain low sensitivity is the fixed way in which the elicitation tasks are set-up. In LT-TTO, the first task always verifies if the utility of a state is equal to (or lower or greater than) -0.5. Such a fixed starting point of the task may reduce sensitivity in the following way. Previous work has shown that respondents may not engage fully in complex TTO tasks [Ramos-Goñi et al, 2018]. As a consequence, respondents may report indifference early in the task, leading to spikes in distribution of values and insensitivity of the observed utility values to severity.

In the present paper, our aim is to test if modifications of cTTO produce data in which correlation between state severity and negative utility is present. Because it may be the sorting question used in cTTO that creates the problem subsequently observed for negative values, we skip the current sorting question altogether in arm B of our design. Because it may be the usual implementation of LT-TTO in which the utility values are left-censored at -1 that diminishes the range of observed values and reduces the amount of information, we continue to elicit the values in arm C after all years have been traded in LT-TTO. Because it may be the fixed bisection procedure used in LT-TTO that results in the distribution of negative utility values having a peak at -0.5 and in an information loss, we change the iterative procedure in arm D by randomizing the tasks presented to the respondent. Arm A used the standard cTTO as a reference. Using various modifications of cTTO in separate arms, we hoped to pinpoint which specific element of cTTO construction reduces its sensitivity.

2 Methods

2.1 Respondents, interviewers, and overall interview design

We recruited respondents from the United Kingdom via Prolific, an online panel of respondents [Palan and Schitter, 2018]. The respondents were interviewed online by six interviewers (graduate students of the Erasmus School of Health Policy & Management, ESHPM, Erasmus University Rotterdam). Throughout the interview, the interviewers shared their screen with the respondents and entered all verbatim responses into the software. All interviews began with the collection of informed consent and the presentation of basic information about the study. The ethical approval for the study was granted by the ESHPM Research Ethics Review Committee.

The interview started with respondents answering basic demographic questions, describing their health using EQ-5D-5L (which includes the EQ VAS visual analogue scale), and describing their experience with health problems. Then, in the main part of the interview, the respondents performed 3 warm-up and 12 actual TTO tasks. Note that every respondent completed only one of four TTO arms (presented in Section 2.2). After the TTO part, the respondents answered questions that aimed to measure their numeracy skills and focus (Section 2.3), the perceived difficulty of the task, and religiosity. The analysis of the impact of numeracy skills and religiosity is beyond the main aim of this paper and will be reported elsewhere.

Information about sample size selection is presented in the Supplementary Materials.

2.2 Study arms

We used four arms, referred to by letters A–D. Arm A used the form of cTTO as in most valuation studies for EQ-5D instruments. It was used as a reference point and also as a means of testing if we can replicate the lack of correlation in our dataset. In the context of the present paper, the following defining characteristics of cTTO are essential. It starts with a comparison of 10 years in health state Q, denoted as $(Q, 10)$, with 10 years in full health, i.e., $(11111, 10)$. If the latter alternative is preferred (a likely outcome), the second comparison is $(Q, 10)$ vs. immediate death, i.e., a sorting question determining whether Q is BTD or WTD. In the former case, the next comparison is $(Q, 10)$ vs. $(11111, T)$ for $T = 5$, and how T is modified in the following tasks depends on the answers. When indifference is reached for $T = T^*$, the linear QALY model implies that $u(Q) = {}^{T^*}\!/_{10}$. In the latter case, LT-TTO starts. Then the respondent is asked to compare $(11111, 10) + (Q, 10)$ (where $+$ stands for *'followed by'*) with $(11111, 10)$, which effectively re-verifies whether a state is WTD [however, substantial framing effects were reported by Jakubczyk et al, 2024]. If the first alternative is preferred (which is usually the case), then the respondent is asked to compare $(11111, 10) + (Q, 10)$ with $(11111, T)$ for $T = 5$, which corresponds to a hypothetical utility $u(Q) = -0.5$. Depending on the answer, the iterative procedure continues by changing T, $0 \leq T \leq 10$. When indifference is reached for $T = T^*$, $u(Q) = {}^{(T^*-10)}\!/_{10}$. Importantly, no $u(Q) < -1$ can be obtained, i.e., utility values are left-censored at -1.

Arms B–D involved modifications to the cTTO. In each, a single element of cTTO was changed ceteris paribus, as described below. Arm B was aimed at removing the effect of the sorting question (i.e., sorting into the WTD domain). Of the two equivalent sorting questions used in arm A in succession, we dropped the one using the comparison vs. immediate death. Hence, in the second question of the TTO task in arm B, LT-TTO was used and the respondent was asked to compare $(11111, 10) + (Q, 10)$ with $(11111, 10)$. If the former was preferred, the state Q is considered BTD, and the task is returned to the regular TTO.

Otherwise, LT-TTO continued.

Arm C was aimed at removing the censoring in -1. After all time T was traded in LT-TTO, the alternative using 11111 only could no longer worsen. In such case, if the respondent chose immediate death over $(11111, 10) + (Q, 10)$, which implies that $u(Q) < -1$, then the time spent in Q was reduced to make the alternative involving Q less dreadful and continue searching for the indifference. The reduction was done in one-year intervals, reduced to half-a-year if direction of preference changed. When the choice between immediate death and $(11111, 10) + (Q, 1)$ was reached, no further changes were possible; hence, the utility values were censored at -10.

Arm D was aimed at increasing the respondents' focus for WTD states by departing from using the same pathway of choice tasks in LT-TTO and instead differentiating the initial choice task corresponding to strictly negative utilities. Instead of always starting with a choice between $(11111, 10) + (Q, 10)$ with $(11111, 5)$ (corresponding to $u(Q) = -0.5$), the first choice task was $(11111, 10) + (Q, 10)$ vs. $(11111, T)$, with T randomly chosen from the set $\{2, 4, 6, 8\}$. Subsequently, the usual rules were applied, i.e., T was changed in 1-year intervals (reduced to 0.5-year intervals after a change in direction).

We used 12 health states in the TTO part (after 3 warm-up TTOs), slightly more than the usual 10 used in the EQ-VT protocol. We decided to add two severe health states to increase the amount of information on WTD states. However, we decided to retain the mild states, to not affect the preferences of the respondents by exposing them to only severe states. The health states were grouped into 20 blocks. The detailed information about how health states were selected and organized into blocks is presented in the Supplementary Materials.

2.3 Analysis

First, in Section 3.1, we describe the characteristics of the respondents, focusing on their demographics, self-assessed health, and experience with health problems. In Section 3.2, we present the descriptive statistics on the distribution of elicited utility values per arm. Section 3.3 is central for the present paper. In that section, we study the association between LSS and utility per study arms, looking at all the health states and at only the WTD or BTD health states [i.e., we use the approach proposed by Gandhi et al, 2019]. In Section 3.4, we perform the analysis at the individual respondent level. We study the inconsistencies within individual respondents by looking at pairs of states that can be Pareto-ranked. For such ranking, we measure the proportion of cases in which the Pareto-dominated was assigned greater or strictly greater utility (two types of analysis). This analysis is done for all states and also after restricting the data to only BTD or only WTD states. We also study the regression coefficients for the utility value by the LSS at the individual respondents level, i.e., the regression was done for each respondent separately. This analysis is done for all states and for WTD states only.

3 Results

3.1 Respondents' characteristics

We collected data from 480 respondents: 256 women, 223 men, and 1 person identifying themselves with different gender. The age ranged between 18 and 77, with a mean of 32.6 and a standard deviation (SD) of 11.

The respondents were mostly in good health. The most prevalent own health states in EQ-5D-5L were: 11111 (31.2%), 11112 (20%), 11121 (12.5%), 11122 (10.8%), 11113 (3.8%), 11123 (3.1%), 11223 (2.3%), with the remaining health states occurring each in fewer than 10 respondents. The mean VAS score amounted to 80.6 with SD equal to 14.7; 95% of respondents reported VAS ≥ 50.

A total of 28.5% respondents reported having experienced serious illness in themselves, 80.4% reported having experienced this in family or friends, 61.5% reported having experienced premature death in family or friends.

3.2 Distribution of utility values

Each block of states was used in 24 interviews, with an exception of two blocks which were used 20 and 28 times, respectively. Looking at individual health states, the number of observations varied because some health states are repeated across multiple blocks in EQ-VT; e.g., 55555 is in every block. Eventually, looking at four ranges of LSS: 6–10, 11–15, 16–20, 21–25, we obtained 912, 1928, 2228, and 692 observations in total, excluding warm-up. Each block was used the same number of times in each arm.

In Table 1, we present the summary statistics for state 55555 and for all states pooled (non-warm-up) split by arm of the study. As can be seen, arm B substantially increased the proportion of WTD states compared to arm A (81.7% vs. 61.7%), while arms C and D did not have an impact in this respect. Intriguingly, at the same time the proportion of utility values $= -1$ of all valuations was reduced to 18.3% for arm B compared to 31.7% for arm A. The mean negative utility was larger (i.e., less negative) for arm B than for arm A. Arm C allowed for utility values < -1, and 10% observations were < -1 (among these, 66% were equal to -10), which substantially decreased the mean valuation for all states and specifically for 55555. Randomizing the negative starting point in arm D slightly decreased the proportion of utility value $= -0.5$ (to 2.5% from 4.2% for arm A), while a higher proportion was observed in arm B (8.3%).

The study was not designed to estimate the disutility coefficients for all dimensions and levels or to produce a complete value set due to the insufficient number of respondents per arm (typically, approx. 1000 respondents participate in valuation studies). Nevertheless, because creating a value set is the ultimate goal of valuation studies, to provide additional information about how the arms in the present study perform in terms of producing a value set, in Supplementary Materials we present the results of such an analysis.

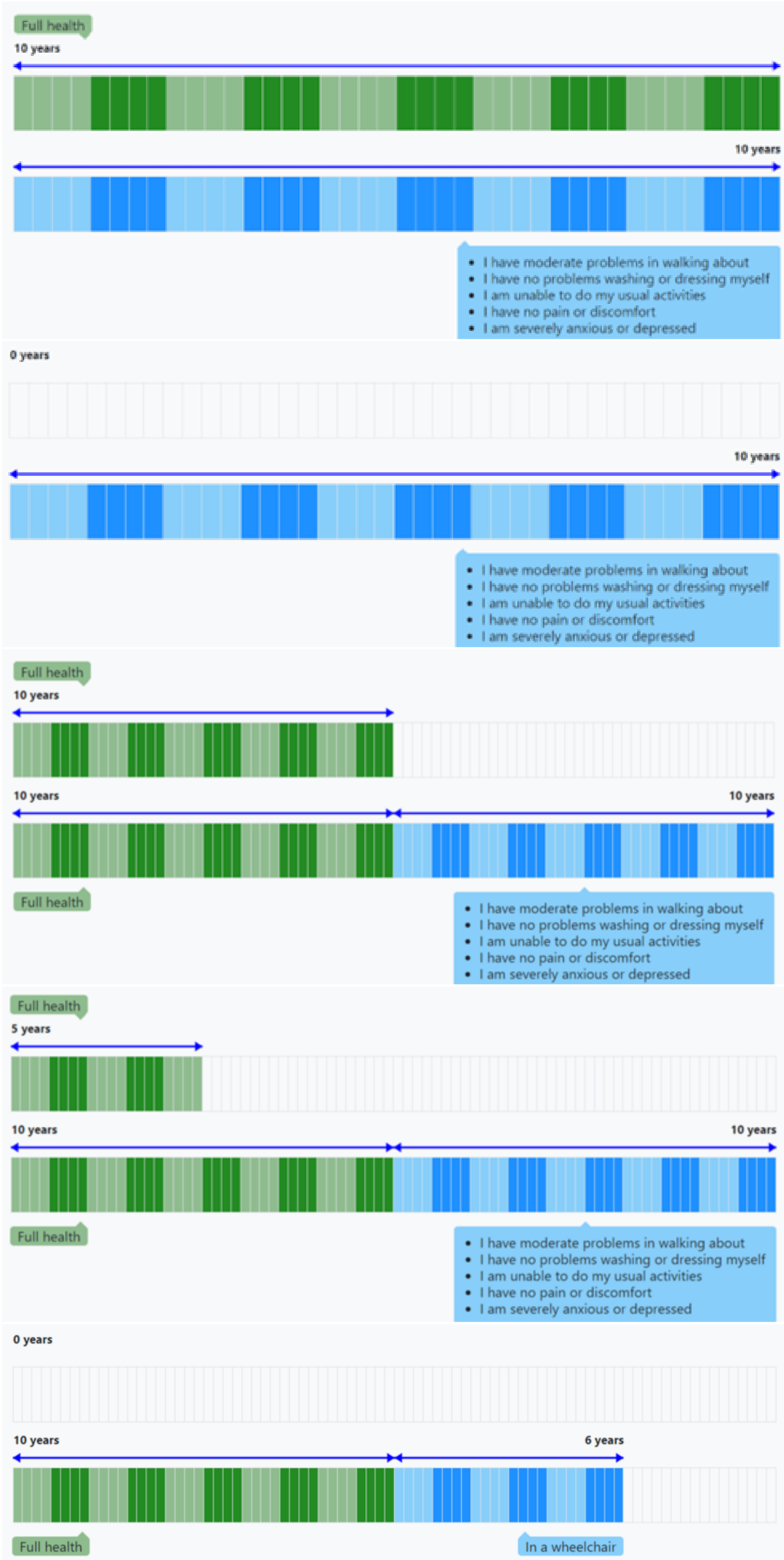


Figure 1: The illustration of regular and lead-time time trade-off (LT-TTO) tasks used in various arms. From top: the starting task (in all arms); the sorting question between better and worse than dead (omitted in arm B); the sorting question based on LT-TTO; the LT-TTO task for the utility value of -0.5; the LT-TTO task for $u < -1$ (arm C only).

Arm	Mean (SD)	Mean (SD) | $u<0$	% $u=1$	$u=0.5$	$u>0$	$u=0$	$u<0$	$u=-0.5$	$u=-1$	$u<-1$
A	−0.406 (0.559)	−0.791 (0.286)	0.8%	3.3%	28.3%	10.0%	61.7%	4.2%	31.7%	n.a.
B	−0.486 (0.495)	−0.677 (0.289)	0.0%	2.5%	15.0%	3.3%	81.7%	8.3%	18.3%	n.a.
C	−2.239 (3.919)	−3.973 (4.375)	0.0%	3.3%	30.8%	10.8%	58.3%	4.2%	0.8%	25%
D	−0.452 (0.545)	−0.776 (0.272)	0.0%	0.8%	26.7%	5.0%	68.3%	2.5%	30.8%	n.a.
A	0.199 (0.644)	−0.752 (0.297)	4.1%	7.6%	68.6%	5.1%	26.3%	2.7%	12.8%	n.a.
B	0.061 (0.656)	−0.571 (0.278)	4.2%	3.2%	48.8%	4.7%	46.5%	7.8%	6.5%	n.a.
C	−0.485 (2.740)	−3.807 (4.186)	4.0%	7.8%	71.9%	4.8%	23.3%	1.9%	0.3%	10.1%
D	0.165 (0.666)	−0.790 (0.246)	2.4%	6.3%	68.3%	2.8%	28.8%	1.3%	12.0%	n.a.

Table 1: Selected descriptive statistics for the utility, u, elicited for state 55555 (top half) and all states pooled (bottom half), split by arm. SD = standard deviation; n.a. = non-applicable because of the arm design.

Arm, subgroup	All states	only $u>0$	only $u<0$
A	−0.068 (0.003, < 0.001; 0.280)	−0.035 (0.001, < 0.001; 0.374)	−0.004 (0.004, 0.237; 0.004)
B	−0.074 (0.003, < 0.001; 0.313)	−0.029 (0.002, < 0.001; 0.330)	−0.018 (0.003, < 0.001; 0.069)
C	−0.168 (0.014, < 0.001; 0.093)	−0.037 (0.001, < 0.001; 0.392)	−0.025 (0.058, 0.671; 0.001)
D	−0.068 (0.003, < 0.001; 0.261)	−0.033 (0.001, < 0.001; 0.356)	0.003 (0.003, 0.262; 0.003)

Table 2: Slopes (standard errors, p-values; and R^2 coefficients) for regressing utility on level sum score (LSS) minus 5 for all states and for subset of states depending on the utility sign.

3.3 Association between negative utility and LSS

We studied how the LSS is associated with the elicited utility values for all states, and then separately for the BTD and WTD states. Because the ways in which zero utility can be assigned differs between arms (specifically, in arm B, no comparison vs. immediate death is used), we decided to slightly modify the usual approach used in Gandhi et al [2019] or Jakubczyk [2023], and split the states based on strict inequality, rather than include $u = 0$ as BTD. What is important is that we maintain the original approach for the WTD states (i.e., states with $u < 0$ are considered as WTD), which is focal for this paper.

The results for all arms are presented in Table 2: the association between the negative utility and LSS was negative in a statistically significant way only for arm B (Fisher Z test Fisher [1925]). Additionally, we present graphically the results in Fig. 2 for arms A (the reference arm), B, and D. The results for arm C are illustrated in Fig. 3 (separately, because of the difference in scale of the ordinate).

In Fig. 4, we present the cumulative distribution functions (CDFs) for the values elicited with individual arms. There seem to be two noticeable differences between arms. First, for arm C values < -1 are observed. Second, the difference between arms A and B appears to be driven by moderate states, i.e., the states for which arm A produces values in the range $[-0.5, 0.5]$, as seen by the two CDFs being separated in this range. Arm B increases the proportion of WTD states and many of these states are assigned only slightly negative utilities, while for arms A, C, and D there are only very few states with utilities in the range $[-0.5, 0]$ and instead a large proportion of states is assigned utility in the range $[0, 0.5]$.

3.4 Individual level analysis

In Table 3, we present the analysis of how often for pairs of states ranked by the Pareto dominance the utilities elicited from a single individual are ordered logically: i.e., we check how often the state dominating in the Pareto sense has a non-strictly or strictly greater elicited utility. For instance, state 34232 dominates state 35245 in the Pareto sense, and it seems warranted to expect that the utility of 34232 should be $\geq$, or $>$ in the strict approach, than the utility of 35245. We analysed such consistence for all states, only for pairs of BTD or of WTD states, and for pairs of states with opposite utility signs (i.e., one state being BTD and the other being WTD).

For the analysis using strict comparison, we additionally accounted for the fact that when both values are censored, no strict relation could be expected (e.g., if the utility equals -1 for both states in arm A, then it should not be treated as inconsistency, as the respondent was unable to express their preference in a more detailed way). The censoring does not impact the result for BTD or opposite-sign states. The proportion is calculated for all ordered pairs of states for all individual respondents.

Table 3: Percentage of responses for which the ordering (non-strict, strict, or strict with allowance for ties when both values censored) of elicited utility values agrees with either the Pareto-ranking (% calculated among all ordered pair of states).

What states	All arms pooled	Arm A	Arm B	Arm C	Arm D
% of correct utility ordering (non-strict)					
utilities > 0	95.1%	94.3%	96.3%	95.9%	94.3%
utilities < 0	89.0%	87.9%	90.0%	88.6%	88.3%
opposite-sign utilities ($\neq 0$)	98.7%	99.1%	97.7%	98.8%	99.3%
all	95.1%	94.8%	94.9%	95.8%	95.0%
% of correct utility ordering (strict)					
utilities > 0	88.8%	88.3%	88.0%	90.4%	87.9%
utilities < 0	52.3%	34.4%	68.3%	49.2%	40.1%
opposite-sign utilities ($\neq 0$)	98.7%	99.1%	97.7%	98.8%	99.3%
all	86.6%	85.2%	87.2%	88.5%	85.6%
% of correct utility ordering (strict, unless both utility values censored)					
utilities < 0	78.7%	81.4%	76.7%	78.3%	80.5%
all	90.4%	90.5%	89.2%	91.2%	90.8%

We regressed the utility values on LSS at the individual respondent level, separately for all states included and only for WTD states. The mean slopes (and SDs) across individuals in each arm were as follows:

- A: all states, $= -0.069$ ($= 0.032$); WTD states, -0.012 ($= 0.030$);
- B: all states, $= -0.073$ ($= 0.029$); WTD states, -0.023 ($= 0.022$);
- C: all states, $= -0.167$ ($= 0.214$); WTD states, -0.187 ($= 0.323$);
- D: all states, $= -0.069$ ($= 0.032$); WTD states, -0.021 ($= 0.047$).

In Figs. S1 and S3 in the Supplementary Materials, we present the distributions of individual-level slopes for all arms using kernel density plots. Additionally, we present in the Supplementary Materials the slopes and the intercepts at the individual respondent level for subgroups of respondents created based on the number of states they considered WTD.

4 Discussion

4.1 Results

In the paper, we tested three modifications of cTTO to verify if these modifications will result in the emergence of correlation between the cTTO-elicited utility and health state severity measured with LSS for WTD states. In our study, we replicated the lack of significant correlation for the standard cTTO (arm A). Of the experimental arms, a statistically significant correlation emerged only in arm B.

In arm B, to sort the health states between WTD or BTD, no comparison vs. immediate death was used. Instead, a single task was used in which both alternatives offered at least 10 years of life, which might have made the sorting task less abhorrent. When designing the study, we hypothesised that the comparison vs. immediate death may be so appalling to some respondents that only very severe states will be considered WTD and subject to LT-TTO. Subsequently, once lead-time is used in LT-TTO, the respondents may avoid living in these severe states by trading off many years in full health, which will result in very negative elicited utility values. Our results in arm B as compared to arm A seem to confirm this hypothesis. First, more states were considered WTD. Second, in the WTD states, the mean utility was less negative. Third, the CDFs for the utility values elicited seem to diverge for the utility values in the range $(-0.5, 0.5)$. The increase of the number of utility values in the range $(-0.5, 0)$ seems to drive the emergence of the correlation between LSS and negative utility.

Our results are in concordance with these reported previously in the literature. Jakubczyk et al [2023] in their arm B used the sorting question just like ours. They reported an increase in the proportion of WTD states compared to the standard cTTO and that a correlation emerged in arm B between the negative utility and other measures of severity. However, in their arm B the TTO implementation for both BTD and WTD states differed substantially from the standard cTTO, and they only used 10 health states in the design. Jakubczyk et al [2024] compared the proportion of WTD states for various sorting questions. Among others, they used the framings that match arm A and arm B in the present paper. Jakubczyk et al [2024] found that when the latter is used instead of the former, the propensity to consider a state WTD increases.

The increased proportion of WTD states in arm B as compared to arm A results in the decrease of the estimated value of the pits state to -0.588 from -0.479. Conveniently, the decrease is not too large, as it is reduced by the increase of mean elicited utilities conditional of a state being WTD. An advantage of arm B over arm A was the reduction of the number of inconsistencies between logically ordered states for WTD states (see Table 3). The proportion of Pareto-ranked states whose utility values were correctly ordered in a strict sense amounted to 68.3% for arm B compared to 34.4% for arm A. Admittedly, this increase is driven by many utility values being censored in -1 in arm A, which results in lack of strict ordering. Nonetheless, such clustering of values in -1 for arm A reduces the amount of information, so the increase in proportion of strict ordering seems to be an advantage.

Experimental arms C and D did not result in the statistically significant correlation between negative utility and LSS. For arm C, a non-significant negative slope was observed (larger than in arm B, in absolute terms), but there was a substantial variation of elicited utility values in the much enlarged range of possible values which resulted in a large estimation error. A much larger sample would be needed to establish the impact of arm C in a more precise manner. Looking beyond the analysis of correlation between LSS and utility, our results in arm C agree with those reported earlier. The proportion of < -1 values among the ≤ -1 values in arm C amounted to approx. 97%, which seems consistent with 92% reported in Jakubczyk et al [2023]. In addition, the mean utility of 55555 elicited in arm C in the present study, -2.239, is close to -2.15 and -2.52 reported in Jakubczyk et al [2023] in two of their study arms (different from our arm C, but also allowing for the elicitation of utility values < -1).

Arm D seems to offer no improvement in the distribution of the utilities obtained.

Finally, note that the study arms did not differ substantially in perceived difficulty (see Supplementary Materials).

4.2 Limitations

We see the following limitations of our study. First, we interviewed respondents from an online panel. Such samples may differ substantially from representative samples of the general population. For instance, in Jakubczyk et al [2024] a much larger proportion of WTD states was observed in an online sample than typically seen in general population. We also used more health states per respondent and a larger proportion of severe states for each respondent than what is common in valuation studies of EQ-5D instruments. In consequence, we would expect to observe substantially fewer WTD observations in a sample obtained using the EQ-VT protocol and coming from a general population, so the assessment of the correlation between LSS and negative utility may require a larger number of respondents. Nevertheless, we see no reason to expect any other impact of using such a sample on the absence or presence of the correlation.

Another limitation of our study is that for simplicity we used no feedback module in any of the arms, i.e., there was no possibility for the respondent to retrospectively indicate some of the utility values as elicited wrongly. In valuation studies using EQ-VT, such a module is used. For instance, in Golicki et al [2019], 8.3% of the cTTO-derived values were flagged by the respondents in the module and removed from subsequent analysis. It would be interesting to see how the proportion would compare between our study arms and what the correlation would look like if only the non-flagged utility values were used. Based on previous studies, whether the flagged observations are used in the modelling seems to have little impact on the value set but a substantial impact on the number of inconsistencies in cTTO values [Wong et al, 2018].

Finally, we acknowledge that studying the regression results when only a subset of all observations is used and the subset is done using the dependent variable, the estimated slopes are driven towards 0 compared to the actual slope in the whole domain of the dependent variable.

4.3 Further research

With regard to the main goal of the paper, the following future research could be considered, as indicated above. First, it would be interesting to see the results for arm C in larger samples. In the literature, utility values < -1 were observed when the elicitation allowed for it [Jakubczyk et al, 2023], so studying the distribution of these values seems warranted. However, samples larger than ours seem needed to obtain results with satisfactory precision.

Second, data for the TTO variant used in arm B could be collected from samples of the general population. Using arm B in the context of valuing paediatric utility instruments such as EQ-5D-Y-3L may be particularly interesting, as the acceptance of immediate death for a child may be even more appalling to the respondents [Lipman et al, 2023; Devlin et al, 2023].

Going beyond the goal of the present paper, we think that our results suggest the following possibly interesting research questions. As presented in Section 3.2, in arm C many observations were censored in -10, which means that the lowering of the censoring threshold did not eradicate censoring but only changed the censoring point. It may indicate that some respondents focus on avoiding living in very severe states even for a relatively short time (e.g. a year) so much that they do not fully internalize the trade-offs [also, see Liao et al, 2023, for attempts to estimate the < -1 utility values]. Qualitative studies may help to understand the actual mechanisms and shed some light on how to interpret very low negative values.

Second, our results demonstrate that changing the sorting question may improve some characteristics of the distribution of elicited utility values. Other sorting questions than those used in our arms A and B are possible, for instance, Jakubczyk et al [2024] used six different framings. Perhaps using some other sorting questions could be embedded in the cTTO and tested for their impact on the elicited values.

5 Conclusion

The observed lack of sensitivity of cTTO-derived data for WTD states seems to result from how the sorting of states into BTD or WTD is done and not from the insensitivity of LT-TTO: LT-TTO in itself is capable of producing values which are sensitive to other measures of health state severity. Replacing the comparison vs. immediate death in cTTO could be considered.

Acknowledgments

The authors thank the EuroQol Research Foundation for funding the present study. However, the views presented here may not be shared by this foundation or the EuroQol Group. We thank the interviewers: Mies Siekerman, Langming Mou, Paul Verhaak, Robert Pattipeilohij, Sinaï Francisco Dias, and Suzan van de Steeg. The authors thank HostLab for providing the software used to conduct the interviews.

References

Bleichrodt H, Wakker P, Johannesson M (1997) Characterizing QALYs by Risk Neutrality. Journal of Risk and Uncertainty 15:107–114

Cokely E, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R (2012) Measuring Risk Literacy: The Berlin Numeracy Test. Judgment and Decision Making 7:25–47

Devlin N, Pan T, Sculpher M, Jit M, Stolk E, Rowen D, van Hout B, Norman R (2023) Using age-specific values for pediatric HRQoL in cost-effectiveness analysis: is there a problem to be solved? If so, how? Pharmacoeconomics 41:1165–1174

Fisher R (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, Scotland

Gandhi M, Rand K, Luo N (2019) Valuation of Health States Considered to Be Worse Than Death-An Analysis of Composite Time Trade-Off Data From 5 EQ-5D-5L Valuation Studies. Value in Health 22:370–376

Golicki D, Jakubczyk M, Graczyk K, Niewada M (2019) Valuation of EQ-5D-5L Health States in Poland: the First EQ-VT-Based Study in Central and Eastern Europe. Pharmacoeconomics 37:1165–1176

Hausman J, Wise D (1977) Social experimentation, truncated distributions, and efficient estimation. Econometrica 45:919–938

Jakubczyk M (2023) Re-revisiting the utilities of health states worse than dead — the problem remains. Medical Decision Making 43:875–885, DOI 10.1177/0272989X231201147

Jakubczyk M, Lipman S, Roudijk B, Norman R, Pullenayegum E, Yang Y, Gu N, Stolk E (2023) Modifying the composite time trade-off method to improve its discriminatory power. Value in Health 26:280–291, DOI 10.1016/j.jval.2022.08.011

Jakubczyk M, Schneider P, Lipman S, Sampson C (2024) This dead or that dead: framing effects in the evaluation of health states. Value in Health 27:95–103

Kennedy-Martin M, Slaap B, Herdman M, van Reenen M, Kennedy-Martin T, Greiner W, Busschbach J, Boye K (2020) Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. The European Journal of Health Economics 21:1245–1257

Liao M, Rand K, Yang Z, Hsu CN, Lin HW, Luo N (2023) Censoring in the time trade-off valuation of worse-than-dead EQ-5D-5L health states: can a time-based willingness-to-accept question be the solution? Quality of Life Research 32:1165–1174

Lipman SA, Zhang L, Shah KK, Attema AE (2023) Time and lexicographic preferences in the valuation of EQ-5D-Y with time trade-off methodology. The European Journal of Health Economics 24:293–305

Miyamoto J, Wakker P, Bleichrodt H, Peters H (1998) The Zero-Condition: A Simplifying Assumption in QALY Measurement and Multiattribute Utility. Management Science 44:839–849

Palan S, Schitter C (2018) Prolific.ac–A subject pool for online experiments. Journal of Behavioral and Experimental Finance 17:22–27

Pickard A, Law E, Jiang R, Pullenayegum E, Shaw J, Xie F, Oppe M, Boye K, Chapman R, Gong C, Balch A, Busschbach J (2019) United States Valuation of EQ-5D-5L Health States Using an International Protocol. Value in Health 22:931–941

Ramos-Goñi J, Craig B, Oppe M, Ramallo-Fariña Y, Pinto-Prades J, Luo N, Rivero-Arias O (2018) Handling Data Quality Issues to Estimate the Spanish EQ-5D-5L Value Set Using a Hybrid Interval Regression Approach. Value in Health 21(5):596–604

Roudijk B, Donders R, Stalmeier P (2022) A threshold explanation for the lack of variation in negative composite time trade-off values. Quality of Life Research 31:2753–2761, DOI https://doi.org/10.1007/s11136-022-03155-6

Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goñi J (2019) Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. Value in Health 22:23–30

Versteegh M, Vermeulen K, Evers S, de Wit G, Prenger R, Stolk E (2016) Dutch Tariff for the Five-Level Version of EQ-5D. Value in Health 19:343–352

von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press

Woloshin S, Schwartz L, Moncur M, Gabriel S, Tosteson A (2001) Assessing Values for Health: Numeracy Matters. Medical Decision Making 21:382–390

Wong E, Ramos-Goñi J, Cheung A, Wong AaAO (2018) Assessing the Use of a Feedback Module to Model EQ-5D-5L Health States Values in Hong Kong. The Patient - Patient-Centered Outcomes Research 11:235–247

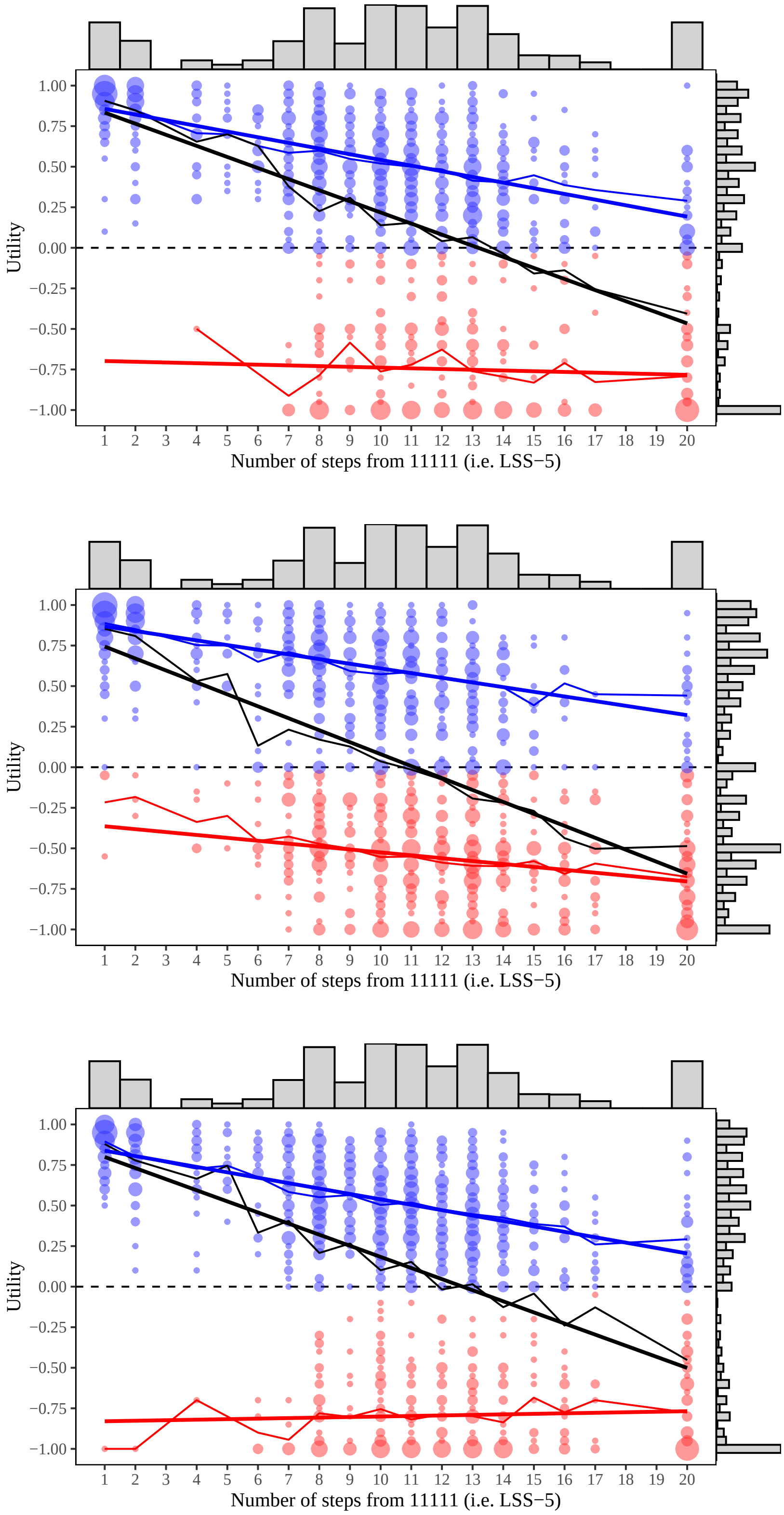


Figure 2: The visualisation of association between level sum score (LSS) and utility, based on the approach of Gandhi et al [2019], for arms A (top), B, and D (bottom). Black, blue, and red lines depict the association for all, strictly positive, and strictly negative utilities, respectively (thick lines for linear regression, thin lines connect mean values).

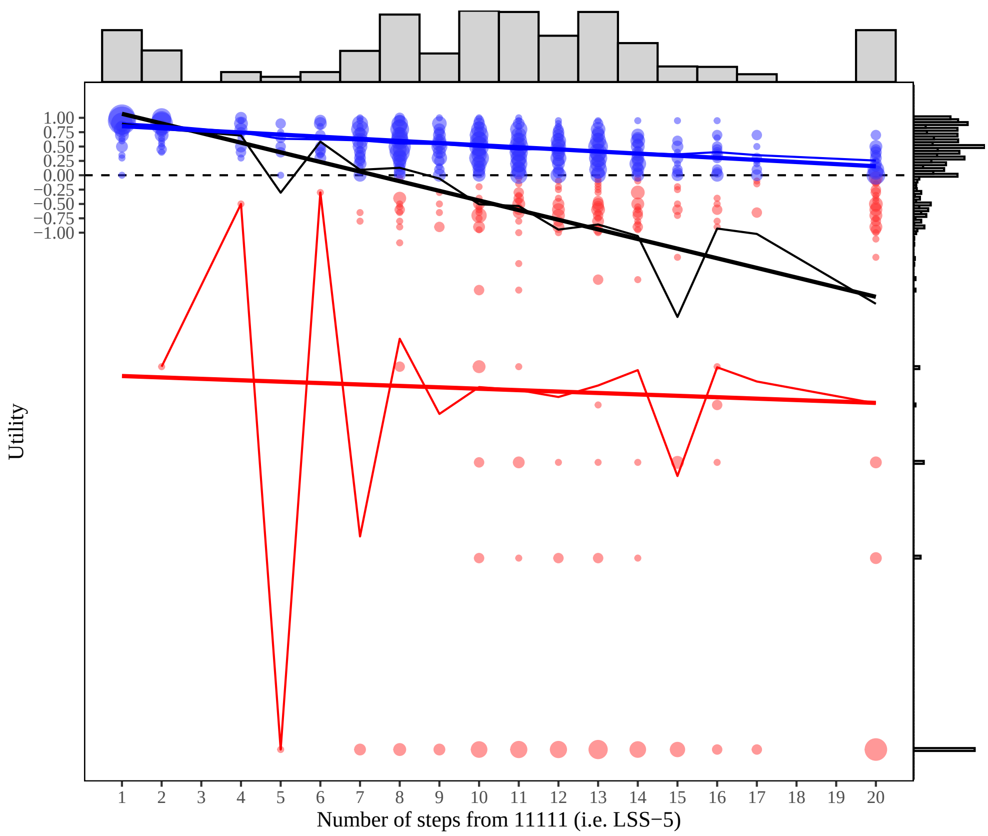


Figure 3: The visualisation of association between level sum score (LSS) and utility, based on the approach of Gandhi et al [2019], for arm C. Black, blue, and red lines depict the association for all, strictly positive, and strictly negative utilities, respectively (thick lines for linear regression, thin lines connect mean values).

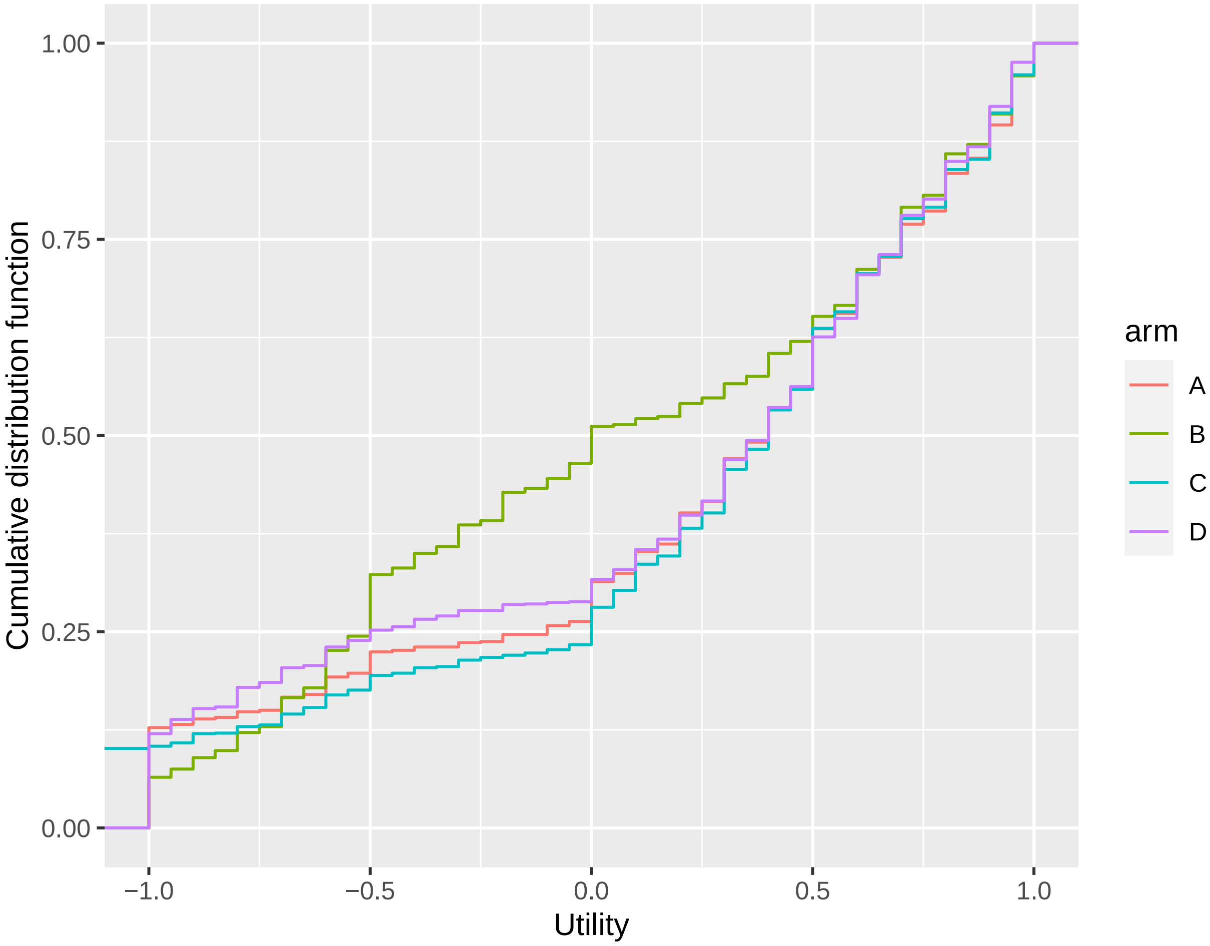


Figure 4: Cumulative distribution functions for utility values elicited with each arm (plot zoomed in to cover only values in the $[-1, 1]$ range.

Supplemental Materials

A. Sample size calculations

The sample size was based on the power calculations with the following assumptions. In the Polish EQ-5D-5L valuation [Golicki et al, 2019], for worse than dead (WTD) observations only, the observed standard deviations (SDs) of the level sum score (LSS) and utility amounted to 4.2 and 0.27, respectively, while the Pearson linear correlation coefficient amounted to -0.06. Detecting an increase of (the absolute value of) correlation by 0.2, i.e., the change of the correlation to -0.26, with a power 75% using a significance level of 0.1 (increased to equate more the probabilities of type I and type II errors) requires 168 observations (i.e., severity-negative utility pairs) per arm. Assuming the proportion of WTD observations about 15% (based on Polish data) and 10 states per person, this number of observations requires approx. 110 respondents per arm.

Eventually, in the project we interviewed 120 respondents per arm, there were 12 tasks per respondent, no observations were lost, and the proportion of WTD observations was larger ($> 20\%$ in all the arms, 46.5% in arm B), which increased the power. In consequence, the actually observed difference of 0.2 in Pearson correlation coefficient between arms A and B was significant with p-value equal to 0.001.

B. Health state selection

The health states were selected as follows. We started with 10 blocks of 10 health states each, exactly as in EQ-VT. Each such block contains a mix of mild, moderate, and severe states; 55555 is in each block. From each block, 4 states with the largest LSS were picked (except for 55555; there was just one tie, which was randomly broken) and a pool of 40 states was created. Each of the 10 blocks was duplicated, and in each copy two empty slots for health states were created. The slots were filled in in the following way. From the pool of states, health states were one by one manually assigned to the blocks while trying to keep the blocks as heterogeneous as possible (technically speaking, by assigning a state to a block for which the minimal Manhattan distance to states already in this block was as large as possible).

The resulting 20 blocks of states were put in a queue. Each block was assigned to one interviewer, and it was used for four consecutive interviews for each of the arms A–D in random order. Such a randomization procedure was used, to make sure that each block was equally used for all arms.

C. Numeracy testing questions

We used two sets of questions to measure the numeracy skills. The first three questions were based on Woloshin et al [2001] as follows:

1. 'What is the most likely number of heads to be obtained in 1000 coin flips using a fair coin?'
2. 'Convert 1% to a proportion, i.e. type how many people out of 1000 it is.'
3. 'Convert the proportion "1 out of 1000" to a percentage.'

with the correct answers being, respectively: 500, 10, 0.1%. We decided to slightly modify the first question. In Woloshin et al [2001], the authors simply asked for the number of heads in 1000 coin flips. We deemed that because the outcome is a random variable, and effectively any answer between 0 and 1000 is possible with non-zero probability, the questions needs to be asked in a more precise way. In Woloshin et al [2001] the authors accepted answers from the symmetrical range encompassing 95% of the probability mass, which seems arbitrary.

The subsequent questions were based on the Berlin Numeracy test [Cokely et al, 2012] with the specific questions being selected in an adaptive way based on previous answers. The details can be found here: `http://www.riskliteracy.org/files/BNT%20Versions.pdf` (last access 15$^{\text{th}}$ Nov 2023).

D. The comparison of value sets produced using data from study arms

We built econometric models, as typically done in valuation studies, to extrapolate the results obtained in the sample to all 3125 EQ-5D-5L health states, split by arm. We used the tobit regression with censoring (at -1 for arms A, B, and D and at -10 for arm C, weighted by observed SD to account for heteroscedasticity). Incremental dummies were used for individual levels, and the dummies were dropped in case of non-intuitive sign of estimated coefficient. We deemed the sample size to be insufficient to interpret individual coefficients per dimensions/level between the arms. Hence, in Table S1, we report the estimated utilities of six states, to allow for comparisons of dimension importance, overall range of utilities, relative importance of levels 2–5, and the proportion of WTD states.

Arms B and C lower the index value for the 55555 health state (the pits state) and they increase the proportion of EQ-5D-5L states that have negative index values. From the individual dimensions perspective, the impact is largest for PD in arm C: worsening this single dimension to level 5 reduces the state value by > 0.8. Looking at the relative importance of levels, arm B makes levels 2–4 relative to level 5 more important, while arm C impacts the results in the opposite way.

Characteristic	Arm A	Arm B	Arm C	Arm D
MO level 5 disutility	0.253	0.315	0.292	0.304
SC level 5 disutility	0.306	0.258	0.327	0.276
UA level 5 disutility	0.233	0.287	0.378	0.300
PD level 5 disutility	0.408	0.456	0.837	0.420
AD level 5 disutility	0.330	0.462	0.319	0.349
$u(55555)$	−0.479	−0.588	−0.931	−0.520
% of states with $u < 0$	18.2%	38.6%	31.3%	21.6%
22222 to 55555 relative disutility	23.7%	29.6%	11.1%	28.8%
33333 to 55555 relative disutility	42.2%	59.6%	22.1%	38.0%
44444 to 55555 relative disutility	86.8%	95.5%	76.4%	88.4%

Table S1: Characteristics of per-arm value sets.

E. Individual-level regression of utility by level sum score for worse-than-dead states

In Fig. S1, we present the distribution of slopes for individual respondents when regressing utility by the level sum score. In Figs. S2 and S3, we present the distribution of the slopes and intercepts when regressing the utility by the level sum score in the subgroups of respondents.

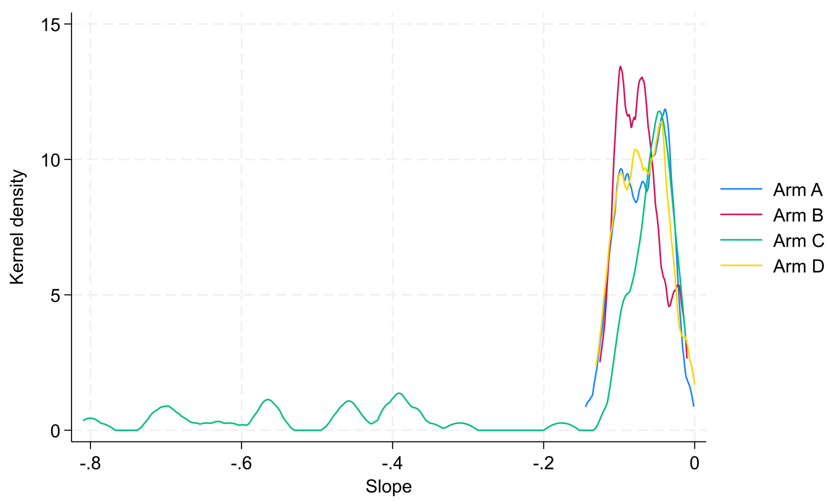
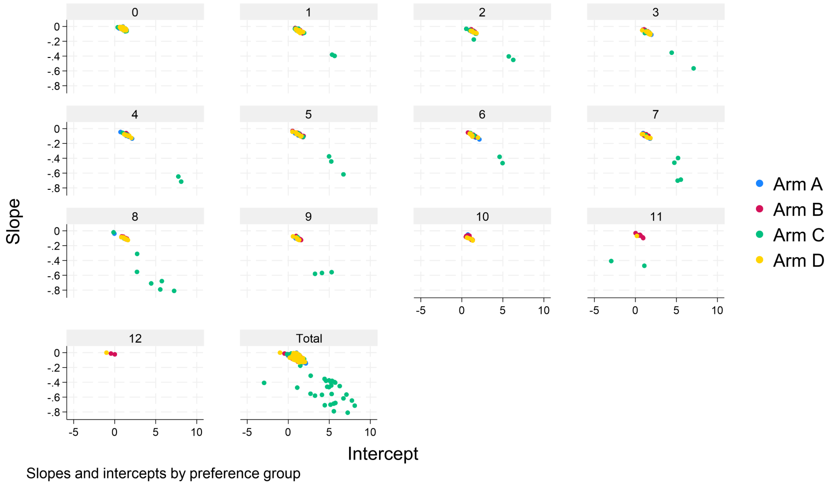


Figure S1: The individual slopes when regressing utility by the level sum score. Visualised per arm using kernel density plots.



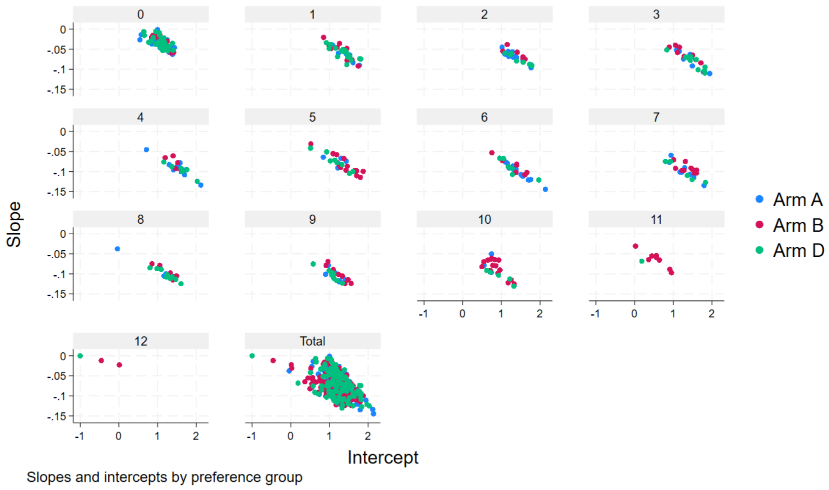


Figure S2: Slopes and intercepts for regressions of utility by the level sum score. Presented per arm, separately for subgroups of respondents depending on the number of states with strictly negative utility. The upper plot for all the arms, the bottom plot zoomed in with arm C removed.

F. Arm perceived difficulty

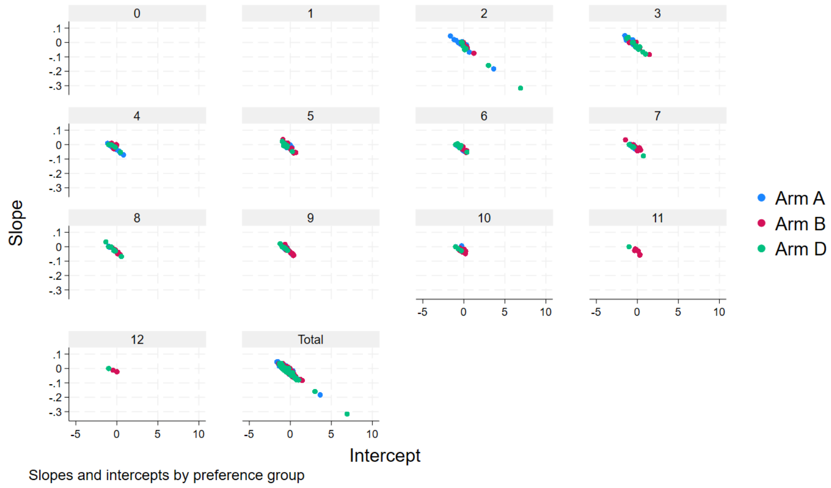


Figure S3: Slopes and intercepts for regressions of utility by the level sum score for WTD states only. Presented per arm, separately for subgroups of respondents depending on the number of states with strictly negative utility.

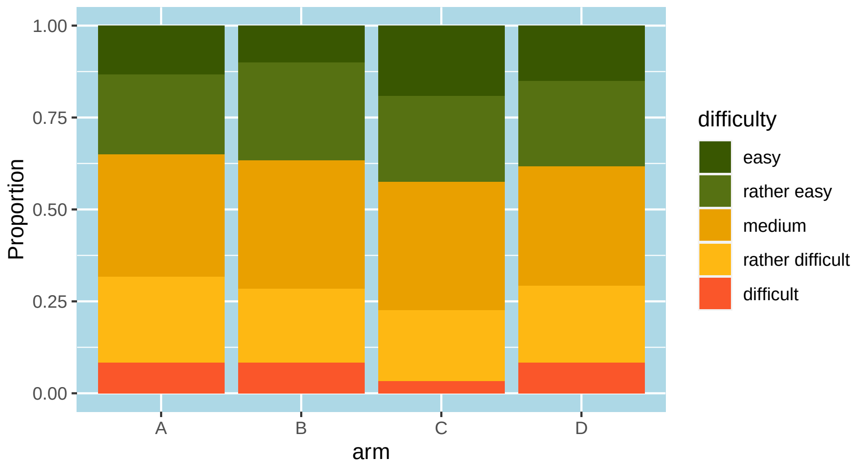


Figure S4: The difficulty of each arm as perceived by the respondents.