A comparison of individual and collective decision making for standard gamble and time trade-off

Arthur Attema, Han Bleichrodt, Olivier L'haridon, Stefan Lipman

QALYs for you and me: A comparison of individual and collective decision making for standard gamble and time trade-off

Arthur E. Attema[a], Han Bleichrodt[b], Olivier l'Haridon[c], Stefan A. Lipman[a]*

[a] Erasmus School of Health Policy & Management, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands.
[b] Erasmus School of Economics, Erasmus University, Rotterdam, The Netherlands
[c] CREM, Université de Rennes 1, Rennes, France
*Corresponding author, E: lipman@eshpm.eur.nl, P: +31.10.4082507.

Key Words: collective decision making, standard gamble, time trade-off.

Abstract

Purpose. Although medical decision making is typically a collective process, Quality-Adjusted Life-Years (QALYs), the preferred outcome measure for cost-utility analyses (CUA), are typically derived from individual preferences over health episodes only. This paper reports the first empirical investigation into the effects of collective decision making on QALY methodology, using both time trade-off (TTO) and standard gamble (SG) tasks.

Methods. We investigated collective decision making in dyads, by means of a mixed-subjects design. Two experimental conditions were used: individual decision making (IDM) and collective decision making (CDM). For subjects in both conditions (n = 163), a baseline measurement for both SG and TTO was obtained for three mild health states, described by means of EQ-5D. Next, subjects completed either a filler task (IDM) or a group measurement (CDM) for the same health states, followed by another individual measurement to determine whether learning effects occurred.

Results. Our data suggested that collective decision-making has little to no effect on: 1) decision quality, and 2) decision outcomes. More specifically, no systematic discrepancies between CDM and IDM were observed in terms of consistency and monotonicity for both methods. Furthermore, SG and TTO utilities remained similar across conditions, and the typical difference in elicited utilities between these methods was not affected.

Conclusions. These findings suggest that consulting with others has little effect on preferences with regard to health outcomes in SG and TTO, although learning effects may occur. This conclusion could be relevant for health state valuation studies, which increasingly utilize personal interview strategies. Additionally, our findings add to the literature of the de-biasing effect of collective decision-making, suggesting that no such effect occurs for SG and TTO.

1. Introduction

Medical decision making, such as when patients decide between different surgical procedures or medical therapies, is typically embedded in a collective process. Although usually individual health outcomes are at stake, decisions about health are frequently made in consultation with significant others, such as spouses, children and medical professionals. Such shared medical decision making is commonly seen as the ideal model of treatment decision making (1). This collective feature of medical decision making is, however, not well-documented within the economic literature on health outcomes research. This line of research deals directly with the valuation of outcomes of treatment decisions, but to our knowledge empirical work comparing individual and collective decision making for outcome measurement is scarce, if not non-existent. That is, Quality-Adjusted Life-Years (QALYs), the preferred outcome measure for cost-utility analyses (CUA), are defined over individual preferences only, without explicit consultation of (significant) others.

The QALYs attributed to health outcomes are obtained by multiplying the duration of the outcomes by quality weights, which represent the health-related quality of life of these outcomes. These quality weights, which are normalized such that 0 represents the subjective weight or value of death and 1 reflects full health, are typically determined through choice-based methodologies (2), such as discrete choice experiments (DCE), standard gamble (SG) or time trade-off (TTO). These health state valuation (HSV) methods are for example used to elicit quality weights for disease-specific health profiles, e.g. dementia (3), colorectal cancer (4) and liver disease (5), but also within more general health state utility frameworks such as EQ-5D or SF-6D (6–11). Perhaps unsurprisingly, as standard health economic theory is relatively silent on collective decision making, these methods are typically applied to the individual case, with subjects in HSV studies deciding about their own (hypothetical) health outcomes (7,12). Compiling such individual preferences for health outcomes from a general public sample enables the estimation of QALYs from the societal perspective, which is, the reference case in CUA (13). Nonetheless, the focus on shared medical decision making in clinical practice (14,15), and the increasing uptake of personal interviews in large HSV studies (as advocated by the EuroQoL Group (7,16)), lead to question how choice-based QALY methodology is affected by moving beyond purely individual decision making. In the present paper, we report the results of a first

empirical investigation into the effects of collective decision making on HSV methodology, specifically on SG and TTO methods used to measure generic health state utilities.

Our focus is on how collective decision making affects QALY weights; i.e. on the effect of deciding collectively on SG and TTO decisions' quality, outcomes and processes. As is well-documented in the health economic literature, QALY weights usually differ between these two methods (17–19). Typically, SG weights, obtained through subjects' decisions between staying in a less-than-perfect health and gambling for full health, are higher than TTO weights, which in turn are derived from the years of less-than-perfect health subjects are willing to trade off to obtain full health. Bleichrodt (20) proposed that the different outcomes produced through these methods can be understood as resulting from inaccurate assumptions with regard to analyzing choices within the SG and TTO method. Conventionally, the difference between SG and TTO was explained as resulting from deviations from the linear QALY model, which has been found to be descriptively inaccurate (21,22). Bleichrodt (20) noted that this explanation is incomplete, since it is based on expected utility (EU) theory, and proposed that the difference between SG and TTO could also result from biases, i.e. descriptive violations of EU theory (as modeled by scale compatibility and prospect theory). Specifically, Bleichrodt (20) postulated that SG will be biased upwards as a result of loss aversion and probability weighting, while TTO is biased upwards due to loss aversion and scale compatibility, and is negatively affected by discounting.

Only recently, empirical work has tested Bleichrodt's (20) predictions, and demonstrated that when most of these biases are measured independently and accounted for, SG and TTO no longer produce different QALY weights (23). However, currently no consensus exists on how these biases are best measured or corrected for. A different strategy could be to instruct individuals to complete SG and TTO in groups, because earlier research using monetary outcomes has documented that collective decision making may have debiasing effects for both risky and delayed outcomes. For example, collective decision making has been associated with less impatience (24), and fewer dynamic inconsistencies (25). Other studies give less firm results, with mixed evidence being reported for risk aversion (26–30), ambiguity aversion (26,31,32) and the violation rate of EU (33–35). Research on household decision making demonstrated that couples' risk attitudes deviate from EU to a lesser extent when couples decide together, although probability weighting is still observed (33). On the other hand, an extensive psychological literature exists suggesting that in some cases detrimental effects of group decision

making can be observed, for example when groups engage in 'groupthink'. In some cases, collective decision making will foster limited information search and enhanced confirmation bias (36,37). As such, under the current state of the literature, it is unknown whether completing SG and TTO in groups will decrease the effect of biases.

Our study adds to the medical decision making literature in several respects. First, we report the first empirical investigation into the debiasing effect of collective decision making on SG and TTO. To this end, we compare (the difference between) SG and TTO estimates between-subjects for groups and individuals, and allude to the debiasing effect of collective decision making. Second, our experiment allows us to disentangle the effect of collective decision making from a mere learning effect. We obtain, for each subject, a baseline measurement for SG and TTO, after which we distinguish between groups and individuals. Groups will engage in collective decision making, while individuals will repeat the SG and TTO measurement individually. As such, we are able to isolate the effect of learning from any difference between collective and individual decision making. Finally, we test whether any possible debiasing effects of collective decision making carry over into a final post-measurement for groups.

The remainder of the paper is organized as follows. Section 2.1 covers our theoretical framework and necessary notational conventions, while sections 2.2 introduces methodology and the experimental procedure. In section 3 the results are presented, whilst section 4 features a discussion of these results and concludes.

2. Methods

2.1. Theoretical framework and notation

In this paper, we only consider chronic health profiles described as (Q, T), with Q denoting health status and T denoting its duration in years. For brevity, we denote immediate death as D and if health status is equal to full health (FH) we write $Q = FH$. Under the assumption of completeness, decision makers are able to form preferences over health profiles, denoted using the conventional notation: $\succ$, $\succcurlyeq$, and $\backsim$ to represent strict preference, weak preference, and indifference, respectively. Most studies applying SG or TTO assume that decision makers form these preferences as modeled within the linear QALY model, i.e.:

$$V(Q,T) = U(Q) * T, \quad \textbf{(1)}$$

Decision makers decide about health profiles, either under certainty (in case of TTO) or under risk (in case of SG). Risk is operationalized by presenting decision maker with lotteries of the following form: $(Q_1, T_1)_p(Q_2, T_2)$, which signifies that health profile (Q_1, T_1) will be realized with probability p, and health profile (Q_2, T_2) with probability $1 - p$.

The SG method involves determining probability p at which decision makers are indifferent between a sure outcome (Q, T_{SG}), and a risky prospect $(FH, T_{SG})_p(D)$. In other words, QALY weights are determined by asking subjects to choose between a number of years (T_{SG}) in health state Q for certain and a gamble with two outcomes, which are FH during the same time period (T_{SG}), and D. Typically, p is varied until the respondent is indifferent between the two alternatives. These SG indifferences are typically evaluated under expected utility (EU) theory (38). The TTO method, on the other hand, asks for a time equivalent in perfect health which yields indifference between (Q, T_{TTO1}) and (FH, T_{TTO2}), with $T_{TTO1} > T_{TTO2}$. In other words, subjects are required to compare T_1 years in health state Q to T_2 years in FH. The number of years T_2 in FH is varied until the respondent is indifferent between the two options.

Given the assumptions listed above, and setting $U(FH) = 1$ & $U(D) = 0$, the SG indifference $(\mathrm{Q}, T_{SG}) \sim (FH, T_{SG})_p(D)$ is evaluated by:

$$U(Q) * T_{SG} = p * (1 * T_{SG}) + (1 - p) * 0, \quad \textbf{(2)}$$

and, thus: $U(Q) = p$.

The TTO indifference $(\mathrm{Q}, T_{TTO1}) \sim (FH, T_{TTO2})$ is evaluated by:

$$U(\mathrm{Q}) * T_{TTO1} = 1 * T_{TTO2}, \quad \textbf{(3)}$$

and, thus, we obtain $U(Q) = T_{TTO2}/T_{TTO1}$.

Bleichrodt (20) proposed that the typical differences between SG and TTO weights may result from biases not accounted for in EU theory or the linear QALY framework, such as discounting, loss aversion and probability weighting. Thus, by evaluating SG and TTO without acknowledging these biasing influences, we should observe a gap between SG and TTO. If collective decision making has debiasing effects on SG and TTO, this gap could decrease, which we test empirically.

2.2. Experiment

Two experimental conditions were used: individual decision making (IDM) and collective decision making (CDM). The main experiment consisted of three parts: Part 1, Part 2 and Part 3, with the experimental conditions IDM and CDM only differing in Part 2. The first part served to establish a baseline measurement for SG and TTO utilities. In the second part, subjects in the CDM condition completed SG and TTO again, whilst discussing amongst each other. Subjects in the IDM condition completed a filler task, which was not related to health states, risk or lotteries, to avoid confounding effects. The questionnaire featured the adaptation by Rohde (39) of Ameriks and colleagues' (40) measure of self-control problems. The results of this filler task are not covered in this paper. In the final part, we established a post-measurement to determine whether learning (IDM) or spillover effects (CDM) occurred, by presenting all subjects with one final repetition of SG and TTO utility elicitation (see Table 1 for an overview of the two conditions). SG and TTO utility weights were obtained by means of a choice list for three health states (see Appendix A and B for instructions and screenshots). The same ordering was used within each part: SG choice lists were completed before TTO choice lists. To test for consistency, a single SG choice list was repeated in Part 1 and Part 3, and also for the collective measurement in CDM.

Table 1. Overview experimental conditions.

		Between-subjects comparisons	
	Condition	IDM (*n = 65)*	CDM (*n = 98)*
Within subjects	Part 1	Individual SG and TTO (I1)	Individual SG and TTO (I1)
	Part 2	Filler task (F)	Collective SG and TTO (G)
	Part 3	Repetition of Individual SG and TTO (I2)	Repetition of Individual SG and TTO (I2)

2.2.1. Sample and procedure

A total of 163 students (78 female) of the Rotterdam School of Management participated in this experiment, with a mean age of 19.37 years (*SD* = 1.57). Experimental sessions lasted for approximately 55 minutes, and subjects were rewarded with course credits for their participation. In total, 98 (49 dyads) participants took part in the CDM condition, and 65 in the IDM condition. The experiment was run on computers in sessions of up to four subjects sitting adjacently in separated cubicles. The experiment was programmed in Matlab, and instructions were provided

on a separate sheet (see Appendix A). An instructor was present at all times to answer any questions subjects might have with regard to the procedure. Subjects were explicitly instructed to refrain from discussing with each other, with the exception of the group part of the experiment in the CDM conditions. In this Part 2, subjects in the CDM condition were seated together at one computer and were instructed to discuss until they arrived at one answer that was satisfactory for both of them. Furthermore, they were told that there were no right or wrong answers and that they should go through the experiment at their own pace. When subjects finished Part 3, several demographics and additional variables were collected.

2.2.2. Health state descriptions

Health state descriptions for SG and TTO were obtained from the EQ-5D-5L classification system (41). The EQ-5D-5L distinguishes between five health domains, i.e., ‘‘mobility’’, ‘‘self-care’’, ‘‘usual activities’’, ‘‘pain/discomfort’’, and ‘‘anxiety/depression’’. Within these domains, this taxonomy uses five health state levels from ‘‘no problems’’ to ‘‘extreme problems/unable to’’. In EQ-5D nomenclature, health states are represented by 5 digit codes like 22113. This example features as a label for a health state with: slight problems (i.e. level 2) with mobility and self-care, no problems with the usual activities and no pain/discomfort (i.e. level 1), and moderate anxiety/depression (i.e. level 3). Four health states were utilized in the SG choice lists and TTO choice lists, one of which was only utilized in the practice list (Q_p: 41321). The remaining three health states reflected an array of mildly aversive health states, in order to avoid health states that could be considered worse than death (42). Additionally, the health states were monotonically increasing in severity, i.e. each consecutive health state featured more severe problems on at least one domain and was identical otherwise. The following health states were used: 11221 (’high’), 21222 (’middle’) and 32322 (’low’), which we denote Q_1, Q_2 and Q_3. In other words, if $T_1 = T_2 = T_3$, assuming monotonicity, we should obtain $(Q_1, T_1) \succ (Q_2, T_2) \succ (Q_3, T_3)$. Subjects completed SG and TTO choice lists for Q_1, Q_2 and Q_3 in the same order for each part of the experiment. To familiarize subjects with the health states in this experiment, before being presented with the choice list elicitation, subjects were required to rate Q1, Q2 and Q3, alongside death on a scale between 0 and 100, where 100 represented full health.

2.2.3. Measurements for SG and TTO

To familiarize subjects with the choice list elicitation, they completed a practice session for both SG and TTO. For choice lists based on the SG method, subjects were faced with a choice between two alternatives. Alternative A would make them certain to live 50 more years in the indicated health state (Q_p, Q_1, Q_2 or Q_3), after which they would die. If they chose Alternative B, they would be taking a gamble. The following instruction was used to clarify the risk of Alternative B: 'On the one hand, you have the chance $(100 \times p\%)$ of living 50 more years (T_{SG}) in full health (i.e. no problems on any dimension), after which you will die, but on the other hand, you have a chance $(100 \times (1-p)\%)$of dying within a week'. Subjects faced choice lists of 10 choices in which Alternative B varied; more specifically, p increased. For each elicitation, a two-pronged approach was used. First, p varied in increments of 10%, between 0% and 100%. After a switching point was obtained at this level, a second choice list was presented, which elicited a probability at the percentage point. For example, if a subject switched at p = 80% in the first choice list, she would face a second choice list that varied between 70% and 80% with increments of 1% (see Appendix B for screenshots).

For choice lists based on the TTO method, Alternative A was the same as for the SG method, i.e. living 50 more years (T_{TTO1}) in the indicated health state (Q_p, Q_1, Q_2 or Q_3), after which they would die. If they choose Alternative B, they would live T_{TTO2} more years in full health (i.e. no problems on any dimension), after which they would die. A similar two-step elicitation procedure was in place, where, in the first choice list, T_{TTO2} varied between 0 and 50 years, with 10 increments of 5 years. In the second choice list, the indifference point of the first list was continued, and a more precise estimate was obtained by presenting subjects with a choice list with 10 increments of 0.5 year. For example, if a subject switched from A to B at T_{TTO2}=35 years, she would face a choice list with Alternative B varying between 30 and 35 with 0.5 year increments (see Appendix B).

3. Results

We present the results of our experiment on the following domains of decision making: a) decision quality, b) decision outcome, and c) decision process (a full transcript of our analyses can be found in the online supplements to this article).

3.1. Data analyses

Each of these decision domains was first analyzed by direct comparisons (i.e. t-tests) at the aggregate level between sessions and conditions. Second, we applied more advanced analyses to the parts on decision quality and decision outcomes, in order to i) determine if collective decision valuation of SG and TTO influences decision making up and above mere learning, and ii) estimate if collective decision making improves subsequent individual decision making. The former approach is referred to as a *'group effect'*, while the latter is referred to as *'carryover effect'*. For the group effect we compared the group answers in the CDM condition (CDM: G) to the repeated individual answers in the control group (IDM: I2). Thus, this comparison consisted of the second time subjects completed SG and TTO utility weights for both conditions, while individuals in CDM completed this second round in groups. To estimate the group effect, we ran generalized linear mixed effect regressions (LMER) with subject random effects and the following fixed effects included: i) learning – dummy indicating whether it concerned a first or repeated session , ii) treatment – IDM or CDM, iii) method – SG or TTO and iv) group – interaction term for learning and treatment. The carryover effect was estimated similarly, where we instead compared CDM: I2 and IDM: I2 to their respective baseline. To estimate this carryover effect, we ran a similar LMER, with the same fixed effects included; i.e., i) learning, ii) treatment, iii) method, and iv) carryover – interaction term for learning and treatment. These analyses were performed with R using the LMER package. For the sake of brevity, we will not present full model statistics for these analyses, but only report fixed effect estimates (FEE) and standard errors (SE) in Table 2.

Table 2. Fixed effect estimates (standard errors) for LMLR analyses for both group and carryover effects	**Decision quality**		**Decision outcome**	
	Consistency	**Monotonicity[a]**	**ΔTariff[b]**	**Δ(SG-TTO)**
Group effect : IDM: I1 vs. I2 | CDM: I1 vs G				
Constant	8.87 (2.02) ***	1.09 (0.65) +	0.13 (0.03) ***	0.06 (0.02) ***
Learning	-1.68 (1.25)	0.64 (0.44)	-0.04 (0.01) ***	0.00 (0.01)
Treatment: CDM	0.15 (2.59)	-2.75 (1.03) **	0.03 (0.03)	0.02 (0.03)
Method: TTO		0.42 (0.28)	0.03 (0.01) ***	
Group: (Learning*Treatment)	-0.74 (1.61)	2.38 (0.86) **	-0.01 (0.01)	-0.01 (0.01)
Health state: middle			-0.04 (0.01) ***	-0.04 (0.01) ***
Health state: high			-0.08 (0.01) ***	-0.08 (0.01) ***
Carryover effect : IDM: I1 vs. I2 | CDM: I1 vs I2				

Constant	8.87 (1.99) ***	1.32 (0.69) +	0.12 (0.03) ***	0.06 (0.02) *
Learning	-1.68 (1.22)	0.67 (0.45)	-0.04 (0.01) ***	0.00 (0.01)
Treatment: CDM	-1.35 (2.30)	-0.51 (0.82)	0.03 (0.03)	0.01 (0.03)
Method: TTO		0.42 (0.26)	0.03 (0.01) ***	
Carryover (Learning*Treatment)	0.76 (1.31)	0.12 (0.55)	-0.01 (0.01)	-0.00 (0.01)
Health state: middle			-0.04 (0.01) ***	-0.03 (0.01) ***
Health state: high			-0.06 (0.01) ***	-0.08 (0.01) ***

Note: *,**, and *** represent significance at $p < 0.05$, 0.01 and 0.001 respectively. $^{+}$ indicates marginal significance at $0.05 < p < 0.10$. [a] binomial regression, [b] difference between each utility weight and its Dutch Tariff for EQ-5D-5L (Versteegh et al., 2016)

3.2. Decision quality

We analyzed decision quality by determining the effect of collective decision-making on our consistency checks and monotonicity of SG and TTO valuations (see Appendix C for results on precision and completion times of SG and TTO).

3.2.1. Consistency

Consistency on repeated SG choices was adequate for all individual tasks (I1 and I2 for both IDM and CDM), with no significant difference between original and repeated elicitation (t-tests, p's>0.07). However, consistency was lower for collective decision making, with significant differences existing between original and repeated decision making (t-test, p<.001). Next, we applied our analytical approach on the absolute difference between original and repeated measurements; hence, we estimated the group effect and carryover effect for consistency (see Table 2). Considering consistency checks were only applied to SG, we drop fixed effects for method in both analyses. We found no significant effects in both our analytical approaches.

3.2.2. Monotonicity

We determined for each subject if utility weights for Q1, Q2 and Q3 were monotonically increasing (i.e. if no violations of monotonicity occurred). A large majority (81% to 100% depending on session) of our subjects assigned monotonically increasing utility to all health states. Next, we applied our approach to estimate the group and carryover effect for monotonicity (see Table 2). Subjects were classified as either violators or non-violators, hence we applied a linear binomial mixed effect model instead of LMER. First, when estimating the group effect, we observe significant effects for: a) treatment and b) group. This indicates that: a) although sampling was random, monotonicity was lower overall for subjects in CDM, and b) monotonicity increased for collective decisions above and beyond learning. No effects of

learning or method were observed. Second, when estimating the carryover effect, we found no significant fixed effects.

3.3. Decision outcome

We analyze decision outcomes using a similar analytical approach, with a focus on both absolute utilities elicited with SG and TTO, and the relative differences between these methods.

3.3.1. Utility weights for SG and TTO

Figure 1 presents the main results on SG and TTO utilities. Several within-subjects trends at the aggregate level can be observed from this figure. First, for many elicitations utility weights appeared to increase after repetition, with significant within-subjects increases for 9 out of 18 subsequent increases (all p's < 0.049). Second, our utility weights for health states Q1, Q2 and Q3 appeared to be lower than the Dutch EQ-5D-5L tariffs for these health states (6), which have been estimated at 0.634, 0.742, and 0.852 respectively. We calculated a difference score between each utility weight and its respective Dutch tariff (denoted ΔTariff), which could be considered a benchmark. We found that these difference scores were significantly larger than 0 for all TTO weights (all p's < 0.033), with the exception of the second repetitions for Q2 (IDM only) and Q3 (both conditions). For SG, we observe utility weights closer to benchmark tariffs. For CDM, SG utility weights at baseline (I1) were significantly lower than the tariff for all health states (all p's <0.001), while for IDM these were also (marginally) significant (all p's <0.08). Subsequent SG utilities (session I2 and G) were often no longer lower than Dutch tariffs for both IDM and CDM, although this did not hold for Q1. These findings indicate that repetition and group decisions appeared to move utility weights closer to the benchmark tariffs, i.e. a trend of increasing utility weights was observed.

Next, we apply our analytical approach and estimate the carryover and group effect on the difference between utility weights and Dutch tariffs, where we ran models with health state included as fixed effect. For both these approaches, we found a significant effect for a) learning, b) method and c) health state dummies. These effects indicate that a) repetition reduces the difference between utility weights and tariff, b) TTO utility weights were more distant from Dutch tariffs and c) the difference between Dutch tariffs and our estimates were increasingly larger for more severe health states. No effect of treatment, group or carryover was observed,

indicating that the positive effect observed on aggregate appears not to be related to collective decisions.

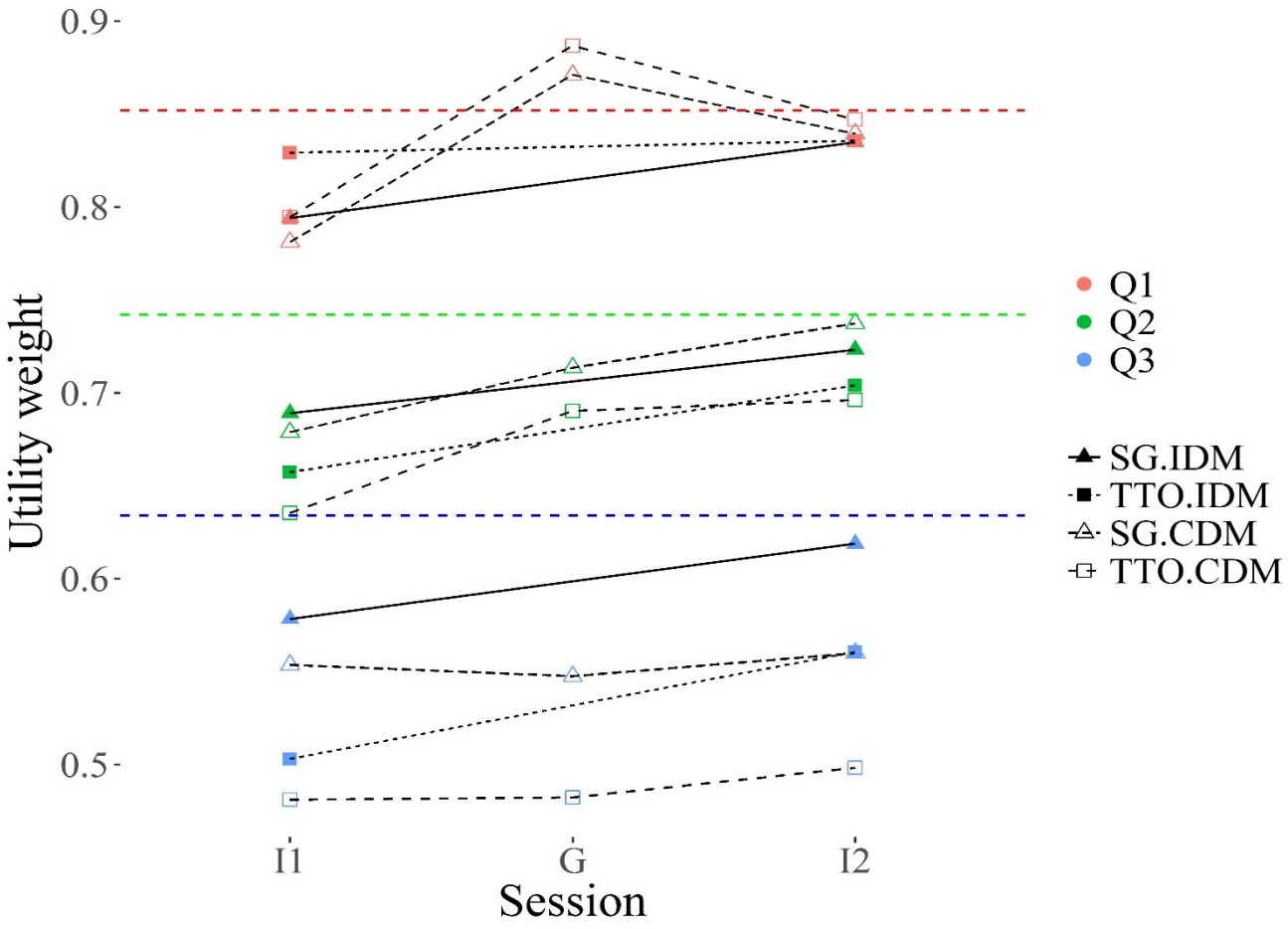


Figure 1: Mean utility weights split by method (SG vs. TTO), session (I1 vs. G vs. I2), health state (Q1 vs. Q2 vs. Q3) and condition (IDM vs. CDM), with colored dashed lines for Dutch tariffs (EQ-5D-5L).

3.3.2. Difference between SG and TTO

Next, we compared the difference between SG and TTO by session and health state (denoted ΔSG-TTO). We found consistent evidence of higher utilities for SG for TTO in health state Q1 (paired t-tests, all p's < 0.011), but no strong evidence for health state Q2 (only significant for CDM-I2, paired t-test, p <0.01) and Q3 (paired t-tests, all p's > 0.11). We found the difference between SG and TTO for baseline measurements (CDM/IDM-I1) pooled across health states to be 0.03 (significantly larger than 0, t-test, p <0.001), suggesting that on average a difference

existed between SG and TTO at baseline. Next, we applied our analytical approach, to estimate group or carryover effects on this difference between SG and TTO (see Table 2). Only fixed effects for health states were significant, indicating that the difference between SG and TTO increased for more severe health states, and no beneficial effects of learning or collective decisions were observed.

3.4. Decision process

Finally, we explored the collective decision-making process by analyzing decision dynamics within dyads completing the CDM task. We estimated to what extent group utility weights deviated from utility weights we observed for the group members at baseline (i.e. I1-CDM). At the aggregate level, a pattern in which the group elicitation falls in-between the two individual estimates is observed most frequently (see Table 4). Such a pattern suggests that a majority of groups reached a consensus somewhere in-between their individual estimates (except for TTO-Q3). Nonetheless, outside consensus group utility weights (lower than min, higher than max) are not uncommon and represent between 28 and 43% of the groups, depending on health state and method. When we investigated within-group consensus (i.e. the proportion of consensus across methods and health states), we observed that groups reach consensus in almost two-thirds of elicitations (64.97%). Only two groups (4%) failed to reach consensus on any elicitation on both SG and TTO. We also found no effect of reaching a consensus or not carrying over into subsequent individual decisions in CDM-I2 (t-tests, all p's > 0.18).

Table 4. Decision process: Location of group utility weight compared to individual weights and median decision weight for high valuators (n = 49).

	SG-Q1	SG-Q2	SG-Q3	TTO-Q1	TTO-Q2	TTO-Q3
Location of utility weight						
Below the min	11	10	4	8	9	3
Above the max	6	10	13	6	12	11
At the min	2	2	4	4	2	4
At the max	2	3	7	2	3	15
In-between	28	24	21	29	23	17
Decision weight	0.43	0.64	0.89	0.44	0.72	1.00

Note: Min and max refer to the lowest and highest individual valuation, respectively.

Next, we estimated the decision weight associated with the highest individual utility in a given group, i.e. the high valuator, for a given decision. We obtained this decision weight by assuming that collective decisions were a weighted summation of individual utilities. In other terms, we calculated decision weight $\propto_H$ of the high valuator in group utility weights (GUW), by rearranging the following equation: $GUW = \propto_H * IUW_H + (1 - \propto_H) * IUW_L$. Here, IUW_H and IUW_L reflect baseline utility weights for the high valuator and their partner who assigned lower utility to that health state, respectively. In this context, if $\propto_H > 0.5$ the high valuator has more weight in decisions, while for $\propto_H < 0.5$ the opposite holds. For the sake of clarity, we removed 6 observations corresponding to the cases where the two individuals' utilities were identical. Table 4 shows that for the best health state (Q3), the group tended to follow the individual with the highest utility, whereas the opposite occurred for the worst health state (both for SG and TTO).

4. Discussion

There is an increasing interest in studies about shared medical decision making, where decisions about health outcomes are arrived at through collective deliberation (1). In this study, we report the first comparison of individual and such collective decision making for health state valuations obtained by SG and TTO. A design was employed in which baseline measurements for both SG and TTO were obtained for three mild health states. Next, either a filler task or a group measurement was completed, followed by another individual measurement to determine whether learning effects, group effects or carryover effects occurred. We analyzed the results of this experiment within three domains of decision making: decision quality, decision outcome, and decision process.

We found no effect of collective decision making with regard to decision outcome, although beneficial effects of learning could be distinguished. We observed a trend of increasing utility weights for SG and TTO, both for collective decisions and for individual decisions. More sophisticated analyses indicated that this increase was related to learning, repetition of SG and TTO (either in groups or individually) increased utility weights, which could be seen as beneficial as this realized a movement towards those of the general population (6). The typical difference between SG and TTO was observed at baseline, although this was less apparent for the least severe health state. Again, a ceiling effect could provide an explanation for this relative

small gap between SG and TTO. Importantly, the gap between SG and TTO was unaffected by collective decision making, and no carryover effects were observed. Finally, we explored decision dynamics within collective decisions. We found that a majority of dyads reached consensus, meaning that SG and TTO utility weights in for their group fell in-between their baseline measurements. For both valuation methods, we observed that the weight the individual with the highest utility decreased with severity. This finding could explain the beneficial effect of collective decisions on monotonicity.

Our results are reassuring for scholars and policy makers who have been applying health state utilities measured at an individual level to medical decision making problems, which in reality often is a collective process. Furthermore, the results indicate that health state valuations can better be improved by adding repetition and practice tasks than by implementing a collective choice task. The latter will be more expensive and burdensome, while generating similar effects as the former.

Collectively, these results add to the evidence base on shared decision making using monetary outcomes. In agreement with the mixed findings of those studies, we do not find a substantial beneficial effect of collective decisions. However, earlier work on collective decisions for monetary choice suggested that groups discount the future less (24,25). Because discounting has a negative effect on TTO values (20), less discounting in the group treatment would cause lower TTO values. Hence, our results suggest that discounting of health outcomes is not affected by collective decision making; an alternative explanation would be that both discounting and loss aversion decrease in group tasks, which would neutralize each other (20). Our results also indicate that collective decision making does not alleviate the typical gap between SG and TTO, which is also partially explained as a result of discounting (20,43). Future research could therefore obtain separate measurements of discounting and loss aversion (and possibly also other traits such as scale compatibility and probability weighting) for health outcomes to test these possibilities.

A drawback of this study was the use of a convenience sample of students, which limits external validity. This was expressed in the lower valuations we observed for TTO compared to those in the general population (6). Still, we feel this first test adds some important insights that can be used in follow-up studies. For one thing, the finding of a substantial learning effect in our student sample suggests that the inclusion of a sufficient number of practice rounds will be

necessary for a less-educated sample representative of the general public. Second, it would be interesting to investigate if our finding of a bias toward the value of the group member with the highest individual utility can be generalized to a more representative sample. Third, future work could replicate our test using married couples, or doctor-patient dyads, who are likely to make real-life medical choices together, increasing the realism of the choice situation.

In sum, a number of conclusions can be drawn from this work. Most importantly, collective decision making does not appear to affect health state valuations compared to individual valuations, above and beyond learning. This is a reassuring result for previous work that has used individually obtained health utilities. Moreover, this suggests that including repetition could have similar beneficial effects as requiring personalized interviews for HSV (e.g. as advocated by EuroQoL in their EuroQoL Valuation Technology protocol (44)).. Second, the preference of the group member with the higher valuation in the individual task gets the highest weight in the group task, which implies that groups tend to behave conservatively regarding the sacrifice of time (TTO) and survival probability (SG). Finally, the difference between SG and TTO does not disappear when moving from an individual to a collective task, which suggests that collective decision making does not help to reduce cognitive biases such as probability weighting. Therefore, other solutions for alleviating these confounding effects, such as more elaborate instructions, practice rounds and correction mechanisms (23) should be considered if one aims to correct for these biases.

References

1. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: What does it mean?(or it takes at least two to tango). Soc Sci Med. 44(5):681–92.
2. Dolan P. The measurement of health-related quality of life for use in resource allocation decisions in health care. In: Culyer AJ, Newhouse JP, editors. Handbook of Health Economics. North Holland: Elsevier; 2000. p. 1723–60.
3. Rowen D, Mulhern B, Banerjee S, Tait R, Watchurst C, Smith SC, et al. Comparison of General Population, Patient, and Carer Utility Values for Dementia Health States. Med

Decis Mak. 35(1):68–80.

4. Fu AZ, Graves KD, Jensen RE, Marshall JL, Formoso M, Potosky AL. Patient preference and decision-making for initiating metastatic colorectal cancer medical treatment. J Cancer Res Clin Oncol. 142(3):699–706.
5. Jo M, Ock M, Lim SY. Estimating Utilities for Liver Diseases Using Standard Gamble Method. Value Heal. 19(7):A838.
6. Versteegh M, Vermeulen K, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch Tariff for the Five-Level Version of EQ-5D. Value Heal. 19(4):343–52.
7. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Econ. 27(1):7–22.
8. Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, et al. A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. Med Care. 54(1):98–105.
9. Kim S-H, Ahn J, Ock M, Shin S, Park J, Luo N, et al. The EQ-5D-5L valuation study in Korea. Qual Life Res. 25(7):1845–52.
10. Kim S-H, Lee S, Jo M-W. Feasibility, comparability, and reliability of the standard gamble compared with the rating scale and time trade-off techniques in Korean population. Qual Life Res. 26(12):3387–97.
11. Brazier J, Ara R, Rowen D, Chevrou-Severac H. A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models. Pharmacoeconomics. 35(1):21–31.
12. McKenna SP, Ratcliffe J, Meads DM, Brazier JE. Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. Health Qual Life Outcomes. 6(1):65.
13. Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, Krahn M, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. Jama. 316(10):1093–103.
14. Makarov D V, Holmes-Rovner M, Rovner DR, Averch T, Barry MJ, Chrouser K, et al. American Urological Association and Society for Medical Decision Making Quality Improvement Summit 2016: Shared Decision Making and Prostate Cancer Screening. Urol Pract.
15. Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making

and patient outcomes. Med Decis Mak. 35(1):114–31.
16. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. Pharmacoeconomics. 34(10):993–1004.
17. Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: Experimental results on the ranking properties of QALYs. J Health Econ. 16(2):155–75.
18. Read JL, Quinn RJ, Berwick DM, Fineberg H V, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. Med Decis Mak. 4(3):315–29.
19. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chronic Dis. 31:697–704.
20. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 456(March):447–56.
21. Wakker P, Deneffe D. Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown. Manage Sci. 42(8):1131–50.
22. Abellán-Perpinán JM, Pinto JL, Méndez-Martinez I, Badia-Llach X. Towards a better QALY model. Health Econ. 15(7):665–76.
23. Lipman SA, Brouwer WBF, Attema AE. QALYs without bias? Non-parametric correction of time trade-off and standard gamble utilities based on prospect theory [Internet]. Working pa. Erasmus University Rotterdam; 2017. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3051140
24. Abdellaoui M, L'Haridon O, Paraschiv C. Do couples discount future consequences less than individuals? Univ Rennes. 1:2013–20.
25. Denant-Boèmont L, Diecidue E, L'Haridon O. Patience and time consistency in collective decisions. Exp Econ. 20(1):181–208.
26. Brunette M, Cabantous L, Couture S. Are individuals more risk and ambiguity averse in a group environment or alone? Results from an experimental study. Theory Decis. 78(3):357–76.
27. Deck C, Lee J, Reyes J, Rosen C. Risk-Taking Behavior: An Experimental Analysis of Individuals and Dyads. South Econ J. 79(2):277–99.
28. Shupp RS, Williams AW. Risk preference differentials of small groups and individuals. Econ J. 118(525):258–83.
29. Ambrus A, Greiner B, Pathak P, others. Group versus individual decision-making: Is there a

shift. Inst Adv Study, Sch Soc Sci Econ Work Pap. 91.
30. Zhang J, Casari M. How groups reach agreement in risky choices: an experiment. Econ Inq. 50(2):502–15.
31. Keck S, Diecidue E, Budescu D V. Group decisions under ambiguity: Convergence to neutrality. J Econ Behav Organ. 103:60–71.
32. Keller LR, Sarin RK, Sounderpandian J. An examination of ambiguity aversion: Are two heads better than one? Judgm Decis Mak. 2:390–7.
33. Abdellaoui M, L'Haridon O, Paraschiv C. Individual vs. couple behavior: an experimental investigation of risk preferences. Theory Decis. 75(2):175–91.
34. Bone J, Hey J, Suckling J. Are groups more (or less) consistent than individuals? J Risk Uncertain. 18(1):63–81.
35. Rockenbach B, Sadrieh A, Mathauschek B. Teams take the better risks. J Econ Behav Organ. 63(3):412–22.
36. Janis IL. Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes. Houghton Mifflin; 1972.
37. Esser JK. Alive and Well after 25 Years: A Review of Groupthink Research. Organ Behav Hum Decis Process. 2(73):116–41.
38. Pliskin JS, Shepard D, Weinstein MC. Utility functions for life years and health status. Oper Res. 28(1):206–24.
39. Rohde KIM. Measuring Decreasing and Increasing Impatience. Manage Sci. Published:mnsc.2017.3015.
40. Ameriks J, Caplin A, Leahy J, Tyler T. Measuring self-control problems. Am Econ Rev. 97(3):966–72.
41. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual life Res. 20(10):1727–36.
42. Dolan P. Aggregating health state valuations. J Health Serv Res Policy. 2(3):160-5; discussion 166-7.
43. Attema AE, Brouwer WBF. The correction of TTO-scores for utility curvature using a risk-free utility elicitation method. J Health Econ. 28(1):234–43.
44. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A Program of Methodological

Research to Arrive at the New International EQ-5D-5L Valuation Protocol. Value Heal. 17(4):445–53.

Appendix A: Example instruction for Part 1

In part 1, you have to perform 2 tasks.

Task 1

Suppose you have to choose between 2 possible life scenarios, which are referred to as Alternative A and Alternative B. In Alternative A, you will be certain to live 50 more years in the indicated health state, after which you will die. For example, suppose the health state is as given below:

Your health state (P):

-You have severe problems in walking about

-You have no problems in washing or dressing yourself

-You have moderate problems doing your usual activities (e.g. work, study, housework, family or leisure activities)

-You have slight pain or discomfort

-You are not anxious or depressed

If you choose Alternative B, you are taking a gamble. On the one hand, you have the chance (X%) of living 50 more years in full health (i.e. no problems on any dimension), after which you will die, but on the other hand, you have a chance (100-X %) of dying within a week.

The task consists of a number of lists of choices between the two alternatives. In every list, Alternative A remains the same, but Alternative B varies.

As you move down the list, Alternative B becomes more attractive, and in some row, you will probably switch from Alternative A to Alternative B. If so, you will also choose Alternative B in all rows below that one, because in these Alternative B is more attractive. Similarly, if you choose Alternative A in a given row, you will also choose Alternative A in all rows above that one, because in these Alternative B is less attractive. The computer takes this into account and automatically selects Alternative B for all rows below the one where you choose Alternative B and Alternative A for all rows above the one where you choose Alternative A.

There are no right or wrong answers, we are only interested in your choices.

You can change your choices as often as you like. Once you are satisfied with your choices, click the "OK" button. Then you can no longer change your choices and you receive the next choice list.

Please now choose the alternative you prefer in each row. If you are ready, you get a prompt on your screen. At that moment, please read the instruction of Task 2 on the next page.

Instructions Task 2

Again, suppose you have to choose between 2 possible life scenarios, which are referred to as Alternative A and Alternative B.

In Alternative A, you will live 50 more years in the indicated health state, after which you will die. For example, suppose the health state is as given below:

Your health state (P):

-You have severe problems in walking about

-You have no problems in washing or dressing yourself

-You have moderate problems doing your usual activities (e.g. work, study, housework, family or leisure activities)

-You have slight pain or discomfort

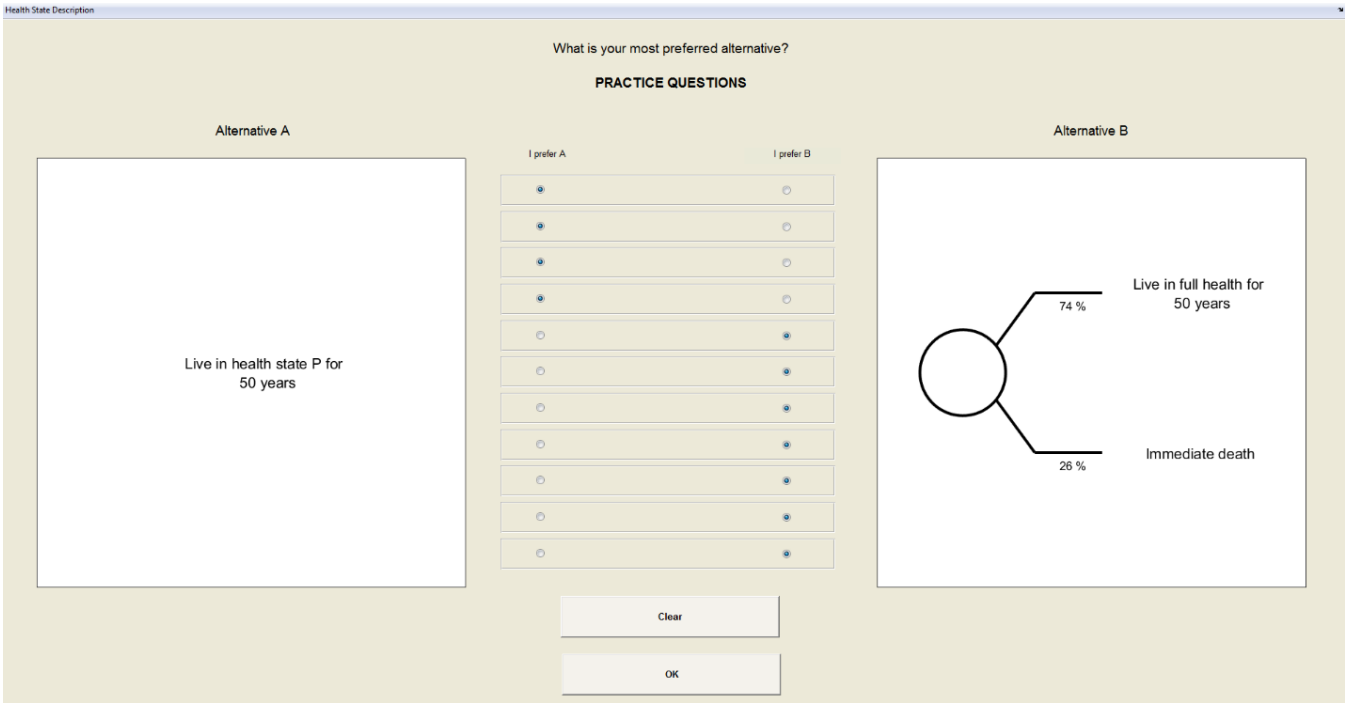
-You are not anxious or depressed

If you choose Alternative B, you will live X more years in full health (i.e. no problems on any dimension), after which you will die.

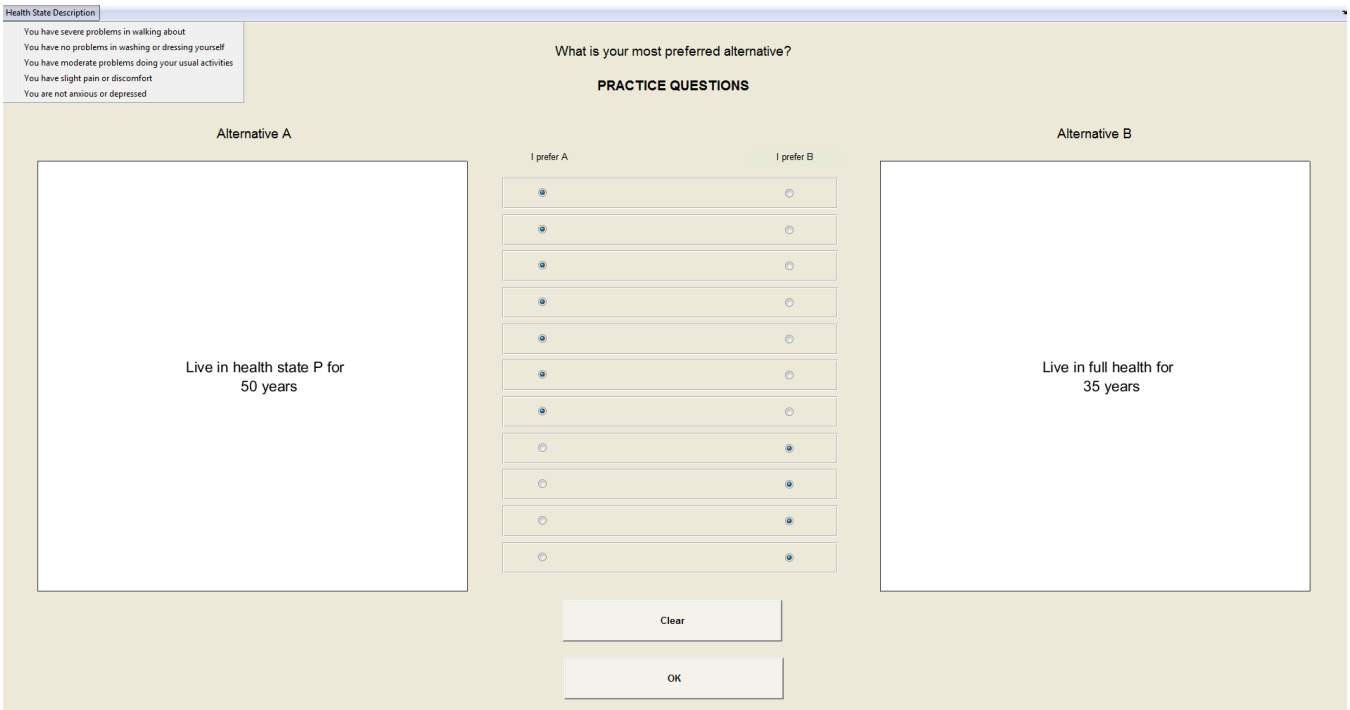
Please choose the alternative you prefer in each row. This procedure is similar as in Task 1.

Appendix B: Screenshots of the experimental program

Task 1: Standard Gamble



Task 2: Time trade-off



Appendix C: Additional results on precision and completion time for SG and TTO

Two additional elements of quality of decision making were analyzed, the precision of utility weights and the completion time for each elicitation. We also estimated the group and carryover effect for these decision elements, which can be found in Table C1.

Table C1. Fixed effect estimates (standard errors) for LMER analyses for both group and carryover effects

	Decision process	
	Precision	**Time**
Group effect : IDM: I1 vs. I2 | CDM: I1 vs G		
Constant	0.03 (0.01) ***	72.74 (3.86) ***
Learning	-0.004 (0.004)	-15.91 (1.99) ***
Treatment	-0.005 (0.009)	-17.37 (4.75) ***
Method: TTO	0.01 (0.002)***	-12.91 (1.25) ***
Group: (Learning*Treatment)	0.009 (0.005) +	15.32 (2.55) ***
Health state: middle		5.54 (1.53) ***
Health state: high		13.55 (1.53) ***
Group effect : IDM: I1 vs. I2 | CDM: I1 vs G		
Constant	0.03 (0.01) ***	76.54 (2.65) ***
Learning	-0.004 (0.004)	-19.89 (1.74) ***
Treatment	-0.001 (0.009)	4.49 (4.11)
Method: TTO	0.01 (0.003) ***	-13.65 (1.09) ***
Carryover (Learning*Treatment)	0.006 (0.005)	-6.53 (2.24) **
Health state: middle		6.34 (1.34) ***
Health state: high		20.35 (1.34) ***

Note: *,**, and *** represent significance at $p < 0.05$, 0.01 and 0.001 respectively. + indicates marginal significance at $0.05 < p < 0.10$.

C.1. Precision

Precision was analyzed both between-subjects and within-subjects. For between-subjects comparisons, we apply Morgan-Pittman tests for equality of variances to compare between session variance within-methods. For example, we compare SG weight variance for state Q1 between session I1 and session I2. These tests indicated the degree to utility weights were heterogeneous between sessions and health states. For IDM variances were not significantly different between I1 and I2 (Morgan-Pittman tests, all p's > 0.16). If we repeat these analyses (I1 vs I2) for CDM, we find a significant decrease (Morgan-Pittman tests, all p's < 0.034) in variance, with the exception of the most severe health state Q3 for both SG and TTO (Morgan-Pittman tests, p's > 0.15). For CDM, we observe significantly smaller variance between the first

individual session and group task (Morgan-Pittman tests, all p's < 0.023). The estimation of fixed group or carryover effects is not possible, as these variance estimates reflect between-subjects heterogeneity. Second, we obtain within-subjects estimates of precision by calculation of variance for utility weights associated with Q1, Q2 and Q3 (see Table C2). These analyses indicate to what extent collective decision-making affected dispersion of utility weights for each individual, i.e. if utility weights elicited in each session become more condensed or dispersed. Next, when we applied our analytical approach to estimate for the group effect and carryover on within-subject variance (see Table C1), we observed only a fixed effect of method, implying higher dispersion for TTO compared to SG. We observed no effects of learning, treatment, group or carryover effects of collective decision making.

Table C2. Decision quality: Mean within-subjects variance and percentages of subjects satisfying monotonicity for each session

	Session 1		Session 2		Session 3
	I1-IDM	I1-CDM	I2-IDM	Group	I2-CDM
Variance for Q1, Q2, & Q3					
SG	0.024	0.030	0.022	0.035	0.034
TTO	0.040	0.043	0.033	0.048	0.043

C.2. Completion time

Completion times were recorded for each session and separately for each health state within each session. Unsurprisingly, for our full sample baseline measurements took longer (5.5 minutes on average) than second individual measurements (little over 3 minutes on average), i.e. repetition decreased time needed for completion ($t(294) = 10.09$, $p < 0.001$). When we focused on subjects in CDM, we observed that group measurements (around 5.5 minutes) took approximately as long as baseline measurement (paired t-test, $t(190) = -0.20$, $p = 0.84$). When applying our analytical approach on within-subjects completion times, similar to our analyses on decision outcomes, fixed effects were also obtained for health states separately, to determine if completion times were affected by severity. In turned out that both when estimating the group and carryover effect almost all fixed effects were significant. The only fixed effect that was not significant was that of treatment in the carryover effects model (p=0.28). Collectively, these findings indicated that decision time consistently decreased: from TTO compared to SG, for repeated sessions, for more

severe health states. Furthermore, the group and carryover effect indicated that collective decisions took longer, while subsequent individual measurements were completed faster for subjects in CDM.