Do Discrete Choice Experiments Approaches Perform Better than Time Trade-off in Eliciting Health State Utilities? Evidence from SF-6Dv2 in China

Shitong Xie, PhD candidate,[1] Jing Wu, PhD,[1*] Xiaoning He, PhD,[1] Gang Chen, PhD,[2] John E. Brazier, PhD [3]

[1] School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China

[2] Centre for Health Economics, Monash Business School, Monash University, Melbourne, Victoria, Australia

[3] School of Health and Related Research, University of Sheffield, Sheffield, UK

[*]Corresponding author: Jing Wu, Ph.D., Professor, E-mail: jingwu@tju.edu.cn, Phone: (086) 15822450465

Funding: This study was funded by the National Natural Science Foundation of China (grant No. 71673197 & No. 71804122).

Abstract

Objectives: To explore the acceptability, consistency, and accuracy of eliciting health state utility values using DCE and DCE with life duration dimension (DCE_{TTO}) as compared with conventional TTO by using the SF-6Dv2.

Methods: During face-to-face interviews, a representative sample of general population in Tianjin, China, completed 8 TTO tasks and 10 DCE/DCE_{TTO} tasks, with the order of TTO and DCE/DCE_{TTO} being randomized. Fixed-effect model and conditional logit models were used for TTO and DCEs data estimation, respectively. Acceptability was assessed by self-reported difficulties in understanding/answering. Consistency was observed by the monotonicity of model coefficients. Accuracy was evaluated by investigating differences between observed and predicted TTO values using intraclass correlation coefficient (ICC), mean absolute difference (MAD) and root mean squared difference (RMSD).

Results: 503 respondents (53.7% males; range from 18-86 years) were included, with comparable characteristics between respondents who completed DCE (N=252) and DCE_{TTO} (N=251). No significant difference was observed in self-reported difficulties among three approaches. The monotonicity of coefficients couldn't be achieved for two DCE approaches even combining the inconsistent levels. The health state utility values generated by DCE were generally higher than that by TTO, whereas DCE_{TTO} lower than TTO. The TTO had a better prediction accuracy than the DCEs.

Conclusions: Two DCE approaches are feasible for eliciting health state utility values; however, they were not considered to be easier to understand/answer than TTO. There are systematic differences in the health state utility values generated by three approaches. The issue of non-monotonicity from two DCE approaches remains a concern.

Key words: Health state utility; Discrete choice experiment; Time trade-off; Acceptability; SF-6D; China.

Highlights:

i. Ordinal approaches such as discrete choice experiments (DCEs) have increasingly been adopted to elicit health state utility values as compared to cardinal approaches like time trade-off (TTO).

ii. However, there were systematic differences in the utility values estimated by the three approaches of TTO, anchored DCE and DCE_{TTO}. The utility values generated by DCE even after mapping were generally higher than that by TTO, whereas DCE_{TTO} lower than TTO. The issue of non-monotonicity from both DCE approaches further remains a concern.

iii. Given health state utility values varied between different elicitation approaches and its important role in healthcare resource allocation, future studies are warranted to identify the most appropriate approach for utility elicitation.

Introduction

Preference-based measures of health-related quality of life (HRQoL), which can be used to generate health state utility values for the calculation of quality-adjusted life years (QALYs), include standardized multi-dimension health state classification system and corresponding country-specific preference weights (also called 'tariffs' or 'value sets') elicited from a representative sample of general population.[1-3] The health state utility values are cardinal values that lie on a 0-1 (death-full health) QALY scale and can include negative values. Examples of the most widely used generic preference-based instruments worldwide are the EQ-5D questionnaire[4] and the Short Form Six-Dimension (SF-6D) questionnaire[5,6].

Health state utility values have been widely elicited using cardinal approaches, such as standard gamble (SG) and time trade-off (TTO).[1,7] However, there are concerns about these approaches because they are likely to be affected by factors other than respondents' preferences for the state, such as time preference and aversion to losses for TTO.[8,9] Furthermore, these approaches are cognitively complex and respondents might have some difficulty in understanding and completing the task, particularly those in vulnerable groups such as the very elderly or children.[9] For these reasons, there has been increasing interest in using ordinal approaches such as discrete choice experiments (DCE), especially for online surveys.[10-12] That is partly because DCE requires respondents to simply choose the one they prefer in the pairwise health states comparisons, instead of going through an iterative process of identifying the indifference point between choices in the TTO approach.[13] On the other hand, it should be noted that when facing choices between two health states that differ in DCEs, there is a large amount of information to process and respondents may struggle to make decisions.

A key problem in using DCE has been how to anchor the values estimated by logit models, i.e. latent utilities, onto the 0-1 QALY scale.[9,14-16] Several studies have attempted to anchor DCE values onto the QALY scale based on external data such as the TTO values of the worst state or the coefficient on the "death" (which was further included as an alternative in the DCE).[9,15] In another variant of DCE, in which an additional dimension of life duration is presented along with the health state, provides a novel alternative to elicit the health state utility values and it requires no separate task or data manipulation for anchoring.[12,13,17-19] Since this approach allowed exploration of the trade-off between quality of life and length of life made by the respondent, the choice task would closely resemble the TTO and the approach is thus referred to as the DCE_{TTO}.[13] A common criticism of TTO that states worse than dead are valued using a different task.[20] A methodological advantage to the DCE_{TTO} is that health states of worse than death can be valued without altering the task.[13,18,19]

While the DCE and DCE_{TTO} appear to be promising approaches for use in future studies, two practical knowledge gaps exist. First, it is still unknown whether the DCE or DCE_{TTO} will be more acceptable to the respondents, compared to the conventional TTO. It has been claimed that DCE tasks are considered simple to complete, and they can be conducted without an interviewer through postal or online survey systems.[9,21,22] So far, no study empirically compared the acceptability and the completion difficulty of these approaches in a single study. Second, there is a lack of head to head comparison of the health state utility values generated by these approaches based on the same instrument in a single study, i.e. whether these approaches could attain similar utility estimates is still unknown. The existing studies have only compared either TTO versus DCE, or TTO versus DCE_{TTO} for condition specific measures or EQ-5D, and they found that different valuation approaches can produce different health state utility values.[9,13,23]

By using the Simplified Chinese version of SF-6Dv2, this study aimed to explore the acceptability, consistency, and accuracy of using DCE and DCE_{TTO} approaches to elicit health state utility values as compared to the conventional TTO approach.

Methods

Instrument

The SF-6Dv2 has six dimensions: physical functioning (PF), role limitation (RL), social functioning (SF), pain (PN), mental health (MH) and vitality (VT). Except for the PN dimension which has six response levels, all others have five levels, with higher values represent more severe states.[1,6] A total of 18,750 (=5*5*5*6*5*5) health states can be defined by the SF-6Dv2 classification system. Detailed description of the SF-6Dv2 can be found elsewhere.[1,6]

Elicitation tasks design

TTO, DCE and DCE_{TTO} elicitation tasks were employed in this study. Appendix Fig 1 in Supplemental Material displays an example of the translated elicitation tasks used in the study.

The composite TTO (cTTO) approach[13,24,25] was used in the TTO task (Appendix Fig 1a), whereby "better than dead (BTD)" and "worse than dead (WTD)" states were valued by conventional TTO and lead-time TTO, respectively.[24,25] In the DCE task (Appendix Fig 1b), respondents were presented with a pair of health states (labelled state A and state B) described by SF-6Dv2, with no reference to the life duration of the states, and asked to indicate which state they preferred. In the alternative DCE_{TTO} task (Appendix Fig 1c), a further dimension describing the number of years the individual would live in that health state followed by death was included. Four levels of life years were chosen: 10, 7, 4, and 1 years.[13] The longest duration was set to 10 years to be commensurate with the standard time frame of the TTO task used in this study.[13,24,25]

Health states selection

The SF-6Dv2 has 18,750 combinations of dimension levels, with more than 175 million potential pairwise combinations generated in the full factorial design. This number would be even more by adding the life duration dimension. Plausibility of combinations of levels of dimensions is also an important consideration. Asking respondents to consider implausible health states is likely to have an impact on the quality of their responses. To balance the statistical efficiency and the respondent's cognitive burden, only one implausible combination (Role Limitations level 1 with Pain level 6) was excluded from the design following previous literature in this study.[19]

Following previous studies, a trade-off was made between the number of health states directly valued and the cognitive burden of respondents.[24,26] In TTO task, 115 health states were valued, including the 6 mildest imperfect health states (211111, 121111, 112111, 111211, 111121, 111112), the worst state (555655), and 108 other states generated based on near orthogonal arrays using SAS® Studio. These 115 health states were split into 18 blocks, all of which contained 1 of 6 mildest health state, the worst state, and 6 block-unique states. The 18 blocks were set for allowing each of the 6 mildest health states to be shown with the same frequency (18/6=3 times per mildest health state). Each respondent was randomly assigned a block for TTO valuation; the order of the appearance of states in each block was randomly allocated.

In both DCE and DCE_{TTO} tasks, 150 pairs of health states (split into 15 blocks) were selected respectively, based on the balanced overlap method. Both main effects and two-way interactions between the levels of each dimension and life years were considered in the experimental design of DCE_{TTO} tasks. The statistical efficiency was maximized with regard to the D-efficiency using Lighthouse Studio 9.6.0 (Sawtooth Software, Inc).[27-29] For the DCE and DCE_{TTO} tasks, each respondent answered 10 pairs of choice tasks with the random block assignment; besides, the task order and the left-right position of health states in each task were also randomized.

Interview design

A face-to-face, computer-based interview was conducted. Two interviewers were involved during the interview with each respondent. According to the study protocol, one of them operated the computer to show all of the questions to the respondent, and the other interviewer recorded problems and difficulties encountered during the interview. Firstly, all respondents were asked to complete the Simplified Chinese version of SF-6Dv2 to be familiar with this classification system.[30] Next, all of the respondents were asked to complete TTO tasks, and half of them randomly selected were assigned to DCE while the rest of them were assigned to

DCE_{TTO} tasks. The order of TTO and DCE/DCE_{TTO} tasks within each respondent were randomized.

Two warm-up questions were used as an example in each task to make sure respondents understood the concept of these tasks before the formal valuation tasks. For TTO, the health states "being in a wheelchair" and "being in a health state worse than dead" were used as examples. For DCE, two stepwise warm-up questions were used. The first warm-up question consisted of a pair of health states described by two random dimensions in SF-6Dv2, and the second one consisted of a pair of states described by adding the other two random dimensions. For DCE_{TTO}, warm-up questions similar to DCE were also used, with the dimension of life duration always added. If the respondents could not understand the warm-up questions, interviewers would keep explaining these questions up to three times. Respondents who still failed to understand the warm-up questions of any of the three tasks were excluded at interview stage.

After the completion of the actual health preference elicitation tasks, respondents were then asked to self-evaluate the difficulties of understanding and answering these tasks based on a 5-level Likert-scale ranging from very easy to very hard. Lastly, respondents' demographic characteristics (age, gender, marital status, ethnic group, household registration), socioeconomic status (education level, employment status, monthly income) and health-related indicators (health insurance coverage, smoking, and alcohol consumption status, presence of chronic conditions) were also collected.

Sample recruitment

A representative sample (target N=500) of the general population were recruited using multi-stage sampling in 11 districts in Tianjin, China, to capture the differences of various geographical regions, population sizes, economic development, and urban-rural proportions. Tianjin city is one of the four municipalities in China, with a total of 16 districts, and more than 15 million permanent population. A quota was set to recruit 45-50 participants in each selected district, stratified with the distributions of age, gender, and education level of the general population in Tianjin.[31,32] Sample recruiting was conducted in publicly accessible places (parks, shops, streets or university campuses) as well as private places (participant's residence) similar to the EQ-5D valuation studies conducted in China.[33,34] Inclusion criteria were that respondents: (1) were 18 years or older; (2) born in mainland China; (3) lived in mainland China for the last five years; (4) were literate and had no disease limiting cognitive function such as dementia; and (5) gave informed consent.

Data collection

A total of 20 interviewers were recruited from Tianjin University and attended a three-day training on the study design, interview protocol, computer software, and interview skills. All interviews were conducted using a laptop computer for displaying questions and recording responses. Data were uploaded and analyzed daily. Very short interviews (less than 4 minutes for any of TTO, DCE or DCE_{TTO} tasks) or logically inconsistent responses (gave same values for all tasks in TTO, always selected the same options as “AAAAA”, or alternately selected the options as “ABABAB” in DCE and DCE_{TTO}) were identified as data with problematic patterns.[35-37] The interviewers were contacted for further confirmation and retraining if necessary. Data with problematic patterns mentioned above were excluded in the final data analysis. Sensitivity analyses were further conducted to explore how these excluded data affected the results reported in the main analysis.

Data analysis

The TTO data were analyzed based on a main-effect model specification (Equation 1):

$$y_i = \alpha + \sum_d \sum_l \beta_{dl} x_{dl} + \varepsilon \quad (1)$$

Where y_i represented the disutility value; α represented the intercept; x_{dl} represented 25 dummy variables indicating the health state described by SF-6Dv2 dimension d at level l, except the first level of each dimension (for reference); β_{dl} represented the coefficient representing the estimated disutility of having problems on dimension d at level l; and ε represented the error term. Considering one respondent completes multiple TTO tasks, in addition to the ordinary least squares (OLS) estimator, the fixed and random effects models were also considered to account for the panel structure in the data.

The DCE data were analyzed under the random utility framework using both a conditional logit model (which assumes a homogenous preference from the respondents) and a mixed logit model (which allows for potential preference heterogeneity among respondents). The utility function consisted of 25 dummy variables similar to what has been shown in Equation 1. The error term was assumed to be independently and identically distributed (iid) with Gumbel distribution. The mixed logit model considers preference heterogeneity by estimating both mean (which represents the average preferences of respondents) and standard deviation. In this study, a SF-6Dv2 dimension was considered as random (with normal distribution) as long as the standard deviation of at least one response level was statistically significant.

A mapping approach was then selected to anchor the latent utility from DCE estimates onto the QALY scale.[9,15,42] Specifically, the latent utility values of the 115 health states directly valuated using the TTO approach were calculated from the DCE estimates. For each of the 115 health states, the mean TTO values were calculated and used as the dependent variable in Equation 2, whilst the predicted latent utility scores served as the independent variable:

$$TTO_i = f(DCE_i) \tag{2}$$

The DCE_{TTO} data was also analyzed under the random utility framework, following the model specification proposed by Bansback et al:[13]

$$U_i = \alpha + \beta t_{dl} + \sum_d \sum_l \lambda_{dl} x_{dl} t_{dl} + \epsilon_i \tag{3}$$

Where U_i represented the latent utility value; t_{dl} represented the life duration, $x_{dl}t_{dl}$ represented the interactions between dimension levels and life duration; t represented the life duration main effect, which was treated as a linear, continuous variable.[13] The DCE_{TTO} value for each health state could be anchored on the QALY scale as:[13,17,19,38,39]

$$V_i = 1 + \frac{\lambda}{\beta} x_{dl} \tag{4}$$

The preferred models for these three valuation approaches were selected based on the following criteria: (1) the monotonicity of the model coefficients, which means that theoretically within each dimension the more severe impairment should have lower values than the milder impairments; (2) the goodness of fit statistics based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC), with lower values indicating better model fit; and (3) the parsimony of the model, meaning that the most parsimonious model would be selected in case two or more models had similar prediction performance. Furthermore, for TTO data, the prediction accuracy could be assessed by comparing predicted and observed mean values for health states valued in the study, using intraclass correlation coefficient (ICC), mean absolute difference (MAD) and root mean squared difference (RMSD). Higher ICC, lower MAD and RMSD values indicated better accuracy. In the main content below, we focused on the results from the preferred models; more details from other estimates can be found in Appendix Tables 9-11.

The comparison of the performance of TTO, DCE and DCE_{TTO} approaches were evaluated in terms of the acceptability, consistency, and accuracy, based on the preferred models. Acceptability was assessed by comparing completion rates, completion time and self-reported difficulties on understanding or answering among these three approaches. Consistency was observed by the monotonicity of model coefficients. The inconsistent coefficients were combined stepwise considering the goodness of fit of model estimation based on AIC and BIC,[5,40] whilst the raw unadjusted results can be found in Appendix Table 2. Based on the preferred model after handling the potential issue of inconsistency, accuracy was evaluated by comparing the predicted health state utility values from each of these three valuation approaches, with the TTO values directly observed from respondents. The ICC, MAD, and RMSD were calculated to assess overall accuracy at predicting observed TTO values.

All statistical analyses were conducted using STATA 14.1. For the comparison of characteristics distributions between subgroups, the t-test was used for continuous variables, while the χ^2 or Fisher exact test was used for categorical variables. Differences in characteristics distributions and the model coefficients are considered statistically significant if the *p*-value <0.05.

Results

Of 576 respondents who were interviewed in July 2018, 73 respondents were excluded because they did not complete the whole interview (N=43) or gave problematic responses (N=30). Finally, a total of 503 respondents were included in this study (Fig 1). The comparison of characteristics between included and excluded respondents is shown in Appendix Table 1. The mean (SD) age of the study sample was 45.4 (16.7) years, ranged from 18-86 years, 53.7% were males. The distributions of characteristics of respondents were close to the Tianjin general population (Table 1). As showed in Table 1, comparable demographic characteristics were observed between the DCE group (N=252) and DCE_{TTO} group (N=251), except only for employment status (p=0.023).

The completion rates were 93.8% for TTO tasks, 95.8% for DCE tasks, and 96.1% for DCE_{TTO} tasks, respectively. While the completion time was significantly shorter for DCE and DCE_{TTO} tasks, no significant difference was observed in self-reported difficulties among the three approaches (Table 2). Sub-group analyses were also conducted for the elderly (aged $\geqslant$ 60 years) and low education level (primary schools or lower) respondents, and showed a consistent result (Appendix Tables 4 and 5). In the DCE group, there was also no significant difference in self-reported difficulties between TTO tasks and DCE tasks (Appendix Table 6), and similar in DCE_{TTO} group (Appendix Table 7).

The fixed-effect model for TTO data and the conditional logit model for both DCE and DCE_{TTO} data were selected for the final data analyses (Appendix Tables 9-11). Table 3 presents the estimated coefficients of the preferred models (i.e. after combination for inconsistent coefficients) on TTO, DCE and DCE_{TTO} data, in which both unanchored and anchored coefficients were reported for DCE and DCE_{TTO}. Most of the coefficients for TTO data were ordered as expected, but levels 4 and 5 in SF dimension, level 3 and 4 in PN dimension and levels 2 and 3 in VT dimension presented slight non-monotonicity. The coefficients for levels 2 and 3 in SF dimension, levels 2 and 3 in VT dimension of DCE, as well as level 2 in RL dimension, and levels 2 and 4 in SF dimension of DCE_{TTO} did not have the expected sign. The combined coefficients were marked with the black squares in Table 3. The goodness of fit was improved after combining the inconsistent levels for all three approaches (Table 3, Appendix Table 2). Furthermore, the sensitivity analysis showed that the excluded data has little impact on the final model results (Appendix Table 3).

The estimated utility values for the 18,750 health states for the SF-6Dv2 based on the TTO, DCE and DCE_{TTO} data are shown in Fig 2. While there was similarity for the very mild states, clear divergence existed in the severe health states. The utility values generated by anchored DCE were generally higher than those by TTO, whereas DCE_{TTO} was lower than TTO. There are 896 health states estimated to be worse than dead using the TTO approach, as compared to

29 and 2400 health states considered to be worse than dead based on DCE and the DCE_{TTO} approaches respectively.

Differences between predicted health state utility values from the three approaches and the observed TTO utility values are reported in Table 4. Since the comparison was against the observed TTO utility values, it is not surprising that the TTO approach had a better prediction accuracy than the DCEs based on all indicators. Comparing the prediction accuracy between DCE and DCE_{TTO} data, it can be seen that overall the DCE data with mapping approach was slightly better than the DCE_{TTO} at predicting TTO values.

Discussion

The key practical issues in using DCE and its variants such as DCE_{TTO} approaches to elicit health state utility values are whether these ordinal approaches will be more acceptable to the respondents and whether they could generate more consistent and accurate health state utility values, as compared to the conventional TTO. To the best of our knowledge, this study provided the first empirical evidence that directly compared the TTO, DCE and DCE_{TTO} approaches in the same study. Furthermore, differing from the previous literature which focused mainly on the respondents in English-speaking developed countries, this study presents the first evidence on the comparison from a non-English speaking country which is also culturally different from western countries.

When compared with the TTO, DCE and DCE_{TTO} were commonly considered to be more acceptable by the respondents in previous studies.[9,17,18,21,22] However, a different finding was found from the respondents in China in this study. Although higher completion rates and shorter completion time were found for DCE and DCE_{TTO} compared with TTO, respondents did not think it was easier to understand or answer the DCE task. This finding was consistent with a previous study that compared TTO and DCE_{TTO} among English-speaking Canadians.[13] Two possible reasons may exist. First, the respondents need to consider two different health states in each DCE or DCE_{TTO} task, while in TTO they only need to consider one health state in each task as the health state of full health is fixed. Second, respondents may struggle more to make choices when the impairment level of two health states in DCEs or DCE_{TTO} tasks vary between each choice task and are often quite similar.

We also found that the proportion of respondents who reported difficulty in answering these three tasks was lower than the previous study.[13] This may be owing to the different interview methods used in these two studies, i.e., the face-to-face interview versus the online survey. During face-to-face interviews, interviewers can clarify respondents' questions during the exercise whilst it is less feasible in an online survey. Consequently, the quality of the data could be better from the face-to-face interview than an online survey.

The results of statistical modelling demonstrated that both the DCE and DCE_{TTO} approaches were feasible to elicit health state utility values. However, although most of the coefficients of the fitted models on these three data sources were logically consistent and statistically significant, it should be noted that several coefficients in RL, VT, and especially in SF dimension, did not have the expected sign. This issue has been reported in previous valuation studies using DCE or DCE_{TTO}. For example, unexpected positive coefficients were observed in urge, urine and coping dimensions of the OAB-5D; concern, breath and pollution dimensions of the AQL-5D using DCE;[9] mobility and self-care dimensions of the EQ-5D-5L;[18] and sad, annoyed and work/housework dimensions of the Child Health Utility 9D (CHU9D) using DCE_{TTO}.[39] However, there was only a very small positive coefficient found for level 3 of VT in the DCE_{TTO} valuation of SF-6Dv2 in the UK.[41] The inconsistency in the estimated coefficients in this study could be due to many factors, such as whether the respondents correctly understood the wordings of the dimension levels, whether they made a rational choice when eliciting their preferences, respondents' cultural and/or educational backgrounds, as well as the choice experiment design. Further studies exploring the issue of inconsistent coefficients in DCE approaches are encouraged.

There were systematic differences in the health state utility values estimated by these three approaches. The utility values generated by DCE were generally higher than that by TTO, whereas DCE_{TTO} was lower than TTO. These differences were also observed in previous studies, which showed that DCE_{TTO} tended to generate lower values, and DCE tended to generate higher values than TTO.[9,13,39] Besides, differences between predicted utility values of these three approaches and observed TTO utility values elicited in this study were similar to a previous study, in which TTO showed a better prediction accuracy than DCE_{TTO}.[9] Nevertheless, it is important to note that the elicited TTO utility values cannot be considered as a "gold standard" with which to compare the values generated from DCE and DCE_{TTO} since these three value sets are derived using different tasks, each requiring different assumptions for econometric modelling techniques.[12,14] The TTO values do, however, provide a benchmark for comparison, and give a relative context to discuss the wider merits and implications of using DCE or DCE_{TTO} as promising alternatives to the TTO.

Several limitations of this study needed to be noted. Firstly, the DCE and DCE_{TTO} approaches were conducted in two separate sub-groups instead of the whole study sample. The trade-off between the design of direct comparison and the cognitive burden of the respondents, which may impact the quality of collected data, must be considered. Among all the characteristics, the only difference found was for the employment status: the DCE sub-group has slightly more respondents in employment than the DCE_{TTO} subgroup (64% vs. 54%). However, when examining their differences in health state valuation using TTO data, a negligible impact on the model estimation was observed (Appendix Table 8). Secondly, considering the relatively small number of health states pairs evaluated given the large descriptive system of the SF-6Dv2, and the limited sample size in this study, there could be an impact on the statistical efficiency of the model estimation. Thirdly, the anchoring method used in this study may affect the utility values generated by DCE data. While several different methods were tried in this study, the mapping method performed the best and all of the other methods showed the same trends when comparing with TTO and DCE_{TTO} data.[42] Furthermore, since the DCE_{TTO} has more dimensions, but in both DCE approaches 150 choice pairs were generated, the design of the DCE_{TTO} tasks will be less efficient as compared to the DCE. Further studies with a larger representative sample and more health state pairs to be evaluated to confirm the properties of DCE and DCE_{TTO} are warranted.

Conclusions

Both DCE and DCE_{TTO} approaches are feasible to elicit health state utility values and generated broadly sensible results. They have higher completion rates and require less completion time than TTO; however, different from most of the previous viewpoints, it is not found to be much easier to understand or answer than the TTO tasks. There exists a systematic difference of the health state utility values predicted by these three approaches, and the issue of non-monotonicity of coefficients from DCE and DCE_{TTO} tasks remains a concern.

References:

1. Brazier, J., Ratcliffe, J., Saloman, J., & Tsuchiya, A. Measuring and valuing health benefits for economic evaluation: OXFORD university press, 2017.
2. Machin, D., & Fayers, P. Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes: Wiley, 2013.
3. Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., & Ganiats, T. G. Cost-effectiveness in health and medicine: Oxford University Press, 2016.
4. EuroQol—a new facility for the measurement of health-related quality of life. *Health policy (Amsterdam, Netherlands)*. 1990; 16: 199-208.
5. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Journal of health economics*. 2002; 21: 271-292.
6. Brazier JE, Mulhern BJ, Bjorner JB, et al. Developing a New Version of the SF-6D Health State Classification System From the SF-36v2: SF-6Dv2. *Medical care*. 2020; 58: 557-65
7. Dolan P. Chapter 32 The measurement of health-related quality of life for use in resource allocation decisions in health care. *Handbook of Health Economics*. 2000; 1: 1723-1760.
8. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health economics*. 2002; 11: 447-456.
9. Brazier J, Rowen D, Yang Y, et al. Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale. *The European journal of health economics*. 2012; 13: 575-587.
10. De Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. *Health economics*. 2012; 21: 145-172.
11. Clark MD, Determann D, Petrou S, et al. Discrete choice experiments in health economics: a review of the literature. *PharmacoEconomics*. 2014; 32: 883-902.
12. Mulhern B, Norman R, Street DJ, et al. One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation. *PharmacoEconomics*. 2019; 37: 27-43.
13. Bansback N, Brazier J, Tsuchiya A, et al. Using a discrete choice experiment to estimate health state utility values. *Journal of health economics*. 2012; 31: 306-318.
14. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Population health metrics. 2003*; 1: 12.
15. Rowen D, Brazier J, Van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. *Medical decision making*. 2015; 35: 328-340.

16. Xie F, Pullenayegum E, Pickard AS, et al. Transforming Latent Utilities to Health Utilities: East Does Not Meet West. *Health economics*. 2017; 26: 1524-1533.

17. Norman R, Cronin P, Viney R. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied health economics and health policy*. 2013; 11: 287-298.

18. Bansback N, Hole AR, Mulhern B, et al. Testing a discrete choice experiment including duration to value health states for large descriptive systems: addressing design and sampling issues. *Social science & medicine*. 2014; 114: 38-48.

19. Norman R, Viney R, Brazier J, et al. Valuing SF-6D Health States Using a Discrete Choice Experiment. *Medical decision making*. 2014; 34: 773-786.

20. Dolan P. Modeling valuations for EuroQol health states. *Medical care*. 1997; 35: 1095-108.

21. Shiroiwa T, Ikeda S, Noto S, et al. Comparison of Value Set Based on DCE and/or TTO Data: Scoring for EQ-5D-5L Health States in Japan. *Value in health*. 2016; 19: 648-54.

22. Devlin NJ, Shah KK, Feng Y, et al. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health economics*. 2018; 27: 7-22.

23. Stolk EA, Oppe M, Scalone L, et al. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. *Value in Health*. 2010; 13: 1005-1013.

24. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in health*. 2014; 17: 445-453.

25. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. *PharmacoEconomics*. 2016; 34: 993-1004.

26. Burgess L, Street DJ, Wasi N. Comparing Designs for Choice Experiments: A Case Study. *Journal of Statistical Theory & Practice*. 2011; 5: 25-46.

27. Chrzan K, Orme B. An overview and comparison of design strategies for choice-based conjoint analysis. *Sawtooth software research paper series*. 2000; 98382.

28. Johnson FR, Lancsar E, Marshall D, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in Health*. 2013; 16: 3-13.

29. Marshall DA, Deal K, Bombard Y, et al. How do women trade-off benefits and risks in chemotherapy treatment decisions based on gene expression profiling for early-stage breast cancer? A discrete choice experiment. *BMJ open*. 2016; 6: e010981.

30. Wu J, Xie S, He X, et al. The Simplified Chinese version of SF-6Dv2: translation, cross-cultural adaptation and preliminary psychometric testing. *Quality of life research*. 2020, 29, 1385–1391. https://doi.org/10.1007/s11136-020-02419-3.

31. National bureau of statistics of China. China Sixth National Census (2010). http://www.stats.gov.cn/ztjc/zdtjgz/zgrkpc/dlcrkpc/ [Accessed February 25, 20202020].

32. Tianjin statistic bureau. Tianjin Statistical Yearbook 2017. http://stats.tj.gov.cn/Item/27612.aspx [Accessed February 25, 202020].

33. Luo N, Liu G, Li M, et al. Estimating an EQ-5D-5L Value Set for China. *Value in health*. 2017; 20: 662-669.

34. Liu GG, Wu H, Li M, et al. Chinese time trade-off values for EQ-5D health states. *Value in health*. 2014; 17: 597-604.

35. Ludwig K, Graf von der Schulenburg JM, Greiner W. German Value Set for the EQ-5D-5L. *PharmacoEconomics*. 2018; 36: 663-674.

36. M. Versteegh M, M. Vermeulen K, M. A. A. Evers S, et al. Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health*. 2016; 19: 343-352.

37. Purba FD, Hunfeld JAM, Iskandarsyah A, et al. The Indonesian EQ-5D-5L Value Set. *PharmacoEconomics*. 2017; 35: 1153-1165.

38. Viney R, Norman R, Brazier J, et al. An Australian discrete choice experiment to value eq-5d health states. *Health economics*. 2014; 23: 729-742.

39. Rowen D, Mulhern B, Stevens K, et al. Estimating a Dutch Value Set for the Pediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration. *Value in health*. 2018; 21: 1234-1242.

40. King MT, Viney R, Simon Pickard A, et al. Australian Utility Weights for the EORTC QLU-C10D, a Multi-Attribute Utility Instrument Derived from the Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30. *PharmacoEconomics*. 2018; 36: 225-238.

41. Mulhern BJ, Bansback N, Norman R, et al. Valuing the SF-6Dv2 Classification System in the United Kingdom Using a Discrete-choice Experiment With Duration. *Medical care*. 2020. https://doi.org/10.1097/MLR.0000000000001324.

42. Xie S, He X, Wu J, et al. PNS240 A comparison of methods for anchoring DCE-derived latent utilities onto the QALY scale: evidence from sf-6d v2. *Value in Health*. 2019; 22: S326-S27.

Table 1 Characteristics of respondents

Characteristics	Total sample (N=503) N (%)	DCE group (N=252) N (%)	DCE_{TTO} group (N=251) N (%)	P-value [a]	Tianjin statistics [b] (%)
Gender [c]				0.445	
Male	270 (53.7%)	131 (52.0%)	139 (55.4%)		54.4%
Female	233 (46.3%)	121 (48.0%)	112 (44.6%)		45.6%
Age (mean [SD])	45.4 (16.7)	45.2 (16.6)	45.6 (16.8)	0.830	NA
Age group (y) [c]				0.934	
18-29	103 (20.5%)	50 (19.8%)	53 (21.2%)		20.0%
30-39	100 (19.9%)	52 (20.6%)	48 (19.1%)		19.9%
40-49	88 (17.5%)	47 (18.7%)	41 (16.3%)		17.7%
50-59	94 (18.7%)	46 (18.3%)	48 (19.1%)		18.8%
≥ 60	118 (23.4%)	57 (22.6%)	61 (24.3%)		23.6%
Education [c]				0.929	
Primary or lower	93 (18.5%)	46 (18.3%)	47 (18.7%)		19.2%
Junior high school	169 (33.6%)	82 (32.5%)	87 (34.7%)		34.6%
Senior high school	115 (22.9%)	58 (23.0%)	57 (22.7%)		22.2%
College or higher	126 (25.0%)	66 (26.2%)	60 (23.9%)		24.0%
Ethnic group				0.668	
Han Chinese	479 (95.2%)	241 (95.6%)	238 (94.8%)		97.4%
Other	24 (4.8%)	11 (4.4%)	13 (5.2%)		2.6%
Household registration				0.653	
Urban	344 (68.4%)	170 (67.5%)	174 (69.3%)		70.0%
Rural	159 (31.6%)	82 (32.5%)	77 (30.7%)		30.0%
Marital status				0.658	
Unmarried	111 (22.1%)	55 (21.8%)	56 (22.3%)		17.1%
Married	352 (69.9%)	176 (69.8%)	176 (70.1%)		75.8%
Divorced	15 (3.0%)	6 (2.4%)	9 (3.6%)		2.0%
Widowed	25 (5.0%)	15 (6.0%)	10 (4.0%)		5.1%
Health insurance					
Urban employee	312 (62.0%)	162 (64.3%)	150 (59.8%)	0.296	NA
Urban & rural resident	182 (36.2%)	87 (34.5%)	95 (37.8%)	0.438	NA
Commercial	93 (18.5%)	47 (18.7%)	46 (18.3%)	0.925	NA
Other	5 (1.0%)	2 (0.8%)	3 (1.2%)	0.686	NA
No	5 (1.0%)	1 (0.4%)	4 (1.6%)	0.216	NA
Employment status				**0.023**	
Employed	297 (59.0%)	162 (64.4%)	135 (53.7%)		NA
Retired	125 (24.9%)	52 (20.6%)	73 (29.1%)		NA
Student	49 (9.7%)	19 (7.5%)	30 (12.0%)		NA
Unemployed	32 (6.4%)	19 (7.5%)	13 (5.2%)		NA
Monthly income (RMB)				0.117	
< 2000	106 (21.0%)	43 (17.1%)	63 (25.1%)		NA
2000-5000	293 (58.3%)	151 (59.9%)	142 (56.6%)		NA
5000-10000	78 (15.5%)	42 (16.7%)	36 (14.3%)		NA
>10000	26 (5.2%)	16 (6.3%)	10 (4.0%)		NA
Smoking status				0.080	
Never	331 (65.8%)	176 (69.8%)	155 (61.8%)		NA

Former smoker	53 (10.5%)	27 (10.7%)	26 (10.4%)		NA
Still	119 (23.7%)	49 (19.5%)	70 (27.9%)		NA
Alcohol consumption				0.135	
Never	277 (55.1%)	146 (57.9%)	131 (52.2%)		NA
Former drinker	53 (10.5%)	20 (7.9%)	33 (13.1%)		NA
Still	173 (34.4%)	86 (34.2%)	87 (34.7%)		NA
Number of chronic conditions [d]				0.331	
0	294 (58.4%)	154 (61.1%)	140 (55.8%)		NA
1	124 (24.7%)	56 (22.2%)	68 (27.1%)		NA
2	44 (8.7%)	22 (8.7%)	22 (8.8%)		NA
3	25 (5.0%)	14 (5.6%)	11 (4.3%)		NA
4 or more	16 (3.2%)	6 (2.4%)	10 (4.0%)		NA

[a] The differences of characteristics distributions between DCE and DCE_{TTO} groups were tested by t-test, chi^2 or Fisher exact tests as appropriate.

[b] All of the data were based on the Tianjin general population. The data of ethnic group was recruited from the Sixth National Census (2010), and other data were recruited from Tianjin Statistical Yearbook 2017; N/A indicates that a direct data was not included in the Yearbook.

[c] The quota sampling was used in which three quotas, i.e., gender, age and education status, were pre-defined on the basis of their distribution in the Tianjin permanent population.

[d] The chronic conditions include: Hypertension, dyslipidemia, diabetes or high blood sugar, cancer or malignant tumor, chronic lung disease, liver disease, heart disease, stroke, kidney disease, stomach or other digestive disease, emotional or psychiatric problems, memory-related disease, arthritis or rheumatism, asthma, or other respondent-reported chronic conditions.

Table 2 The acceptability of TTO, DCE and DCE_{TTO} tasks

Characteristics Mean (SD) / N (%)	**TTO tasks (N=503)**	**DCE tasks (N=252)**	**DCE_{TTO} tasks (N=251)**	**P-value (TTO vs. DCE)**	**P-value (TTO vs. DCE_{TTO})**	**P-value (DCE vs. DCE_{TTO})**
Completion time (min)	12.8 (7.1)	8.9 (4.5)	8.5 (5.6)	**<0.001**	**<0.001**	0.354
Self-reported difficulty level of understanding				0.919	0.295	0.184
Very easy	63 (12.5%)	26 (10.3%)	33 (13.1%)			
Easy	254 (50.5%)	127 (50.4%)	139 (55.4%)			
Moderate	148 (29.4%)	79 (31.3%)	63 (25.1%)			
Hard	32 (6.4%)	17 (6.7%)	16 (6.4%)			
Very hard	6 (1.2%)	3 (1.3%)	0 (0%)			
Self-reported difficulty level of answering				0.360	0.602	0.052
Very easy	55 (11.0%)	17 (6.7%)	34 (13.5%)			
Easy	218 (43.3%)	107 (42.5%)	115 (45.8%)			
Moderate	155 (30.8%)	87 (34.5%)	73 (29.1%)			
Hard	59 (11.7%)	34 (13.5%)	23 (9.2%)			
Very hard	16 (3.2%)	7 (2.8%)	6 (2.4%)			

[1] The differences of completion time between groups were tested by t-test; the differences of distributions of self-reported difficulty level of understanding or answering were tested by chi^2 test.

Table 3 Adjusted estimated coefficients of the fitted models

	TTO data (N=503)		DCE data (N=252)			DCE_{TTO} data (N=251)		
	Fixed effects OLS model		Conditional logit model (Latent utility)		Anchored with Mapping	Conditional logit model [a] (Latent utility)		Anchored with coef. of life duration (coef. = 0.384)
	Coef.	SE	Coef.	SE	Coef.	Coef.	SE	Coef.
Physical functioning								
PF2	-0.032	0.023	-0.175	0.106	-0.036	-0.022	0.014	-0.056
PF3	-0.040	0.024	-0.259	0.101	-0.053	-0.022	0.014	-0.056
PF4	-0.136***	0.022	-0.422***	0.108	-0.086	-0.090***	0.018	-0.234
PF5	-0.410***	0.022	-1.795***	0.131	-0.364	-0.169***	0.017	-0.441
Role limitation					--			
RL2	-0.036	0.021	-0.046	0.106	-0.009	0.000	--	0.000
RL3	-0.052*	0.023	-0.144	0.105	-0.029	-0.020	0.018	-0.052
RL4	-0.065**	0.023	-0.202	0.104	-0.041	-0.038*	0.019	-0.099
RL5	-0.086***	0.023	-0.540***	0.116	-0.110	-0.043**	0.017	-0.113
Social functioning					--			
SF2	-0.110***	0.021	**0.252****	0.088	**0.051**	**0.088*****	0.018	**0.229**
SF3	-0.112***	0.022	**0.338****	0.113	**0.069**	**-0.005**	0.017	**-0.013**
SF4	-0.125***	0.019	-0.255	0.108	-0.052	**0.036***	0.015	**0.093**
SF5	-0.125***	0.019	-0.332**	0.109	-0.067	**-0.022**	0.018	**-0.058**
Pain								
PN2	-0.081***	0.023	-0.028	0.082	-0.006	-0.029	0.020	-0.075
PN3	-0.082***	0.020	-0.028	0.082	-0.006	-0.034	0.019	-0.087
PN4	-0.082***	0.020	-0.028	0.082	-0.006	-0.060**	0.019	-0.157
PN5	-0.333***	0.024	-1.309***	0.128	-0.266	-0.167***	0.020	-0.436
PN6	-0.350***	0.024	-1.689***	0.143	-0.343	-0.199***	0.021	-0.518
Mental health					--			
MH2	-0.037	0.021	-0.041	0.112	-0.008	-0.047**	0.016	-0.123
MH3	-0.118***	0.024	-0.215	0.113	-0.044	-0.047**	0.016	-0.123
MH4	-0.122***	0.022	-0.671***	0.100	-0.136	-0.058***	0.016	-0.152
MH5	-0.135***	0.022	-0.671***	0.100	-0.136	-0.135***	0.020	-0.353
Vitality								
VT2	-0.065***	0.019	**0.289***	0.114	**0.059**	-0.001	0.017	-0.003
VT3	-0.065***	0.019	**0.106**	0.106	**0.022**	-0.033*	0.016	-0.086
VT4	-0.114***	0.022	-0.226*	0.102	-0.046	-0.086***	0.016	-0.224
VT5	-0.123***	0.023	-0.420***	0.105	-0.085	-0.093***	0.018	-0.243
Log likelihood	-1579.251		-2467.7970			-2634.6203		
AIC	3204.5030		4979.5930			5217.2410		
BIC	3349.4040		5123.1290			5473.7490		

[a] The coefficients for DCE_{TTO} data were the interactions between dimension levels and life duration, for example, the PF2*life duration. The coefficient of life duration is 0.384 (p <0.001), with the SE of 0.032.

Note: The coefficients in bold meant non-monotonic with opposite sign. The coefficients in square meant non-monotonic while adjusted by combining the non-monotonic levels, which meant the combined levels had the same disutility from the reference level (i.e. the first level) in each dimension. Levels 2 to 3 of PF were combined which contains limited a little in vigorous activities to moderate activities. Levels 1 to 2 of RL were combined which contains accomplish less than you would like none of time to a little of time. Levels 2 to 3 of SF/MH/VT were combined which contains social activities are limited/depressed or very nervous/worn out a little of time to some of time. And Levels 2 to 4 of "Pain" were combined which contains very mild pain to severe pain. *** p <0.001, ** p <0.01, * p <0.05. AIC, Akaike information criterion; BIC, Bayesian information criterion.

Table 4 The accuracy of three approaches compared to the observed TTO data

	TTO data (N=503)	DCE data (N=252)	DCE_{TTO} data (N=251)
ICC	0.938	0.872	0.873
No. (%) of differences >0.05 from observed TTO	32 (27.8%)	23 (20.0%)	23 (20.0%)
No. (%) of differences >0.1 from observed TTO	47 (40.9%)	61 (53.0%)	62 (53.9%)
MAD from observed TTO	0.1003	0.1339	0.1620
RMSD from observed TTO	0.1311	0.1710	0.2154

ICC, intraclass correlation coefficient; MAD, mean absolute difference; RMSD, root mean squared difference.
Higher ICC, lower MAD and RMSD indicated better accuracy.

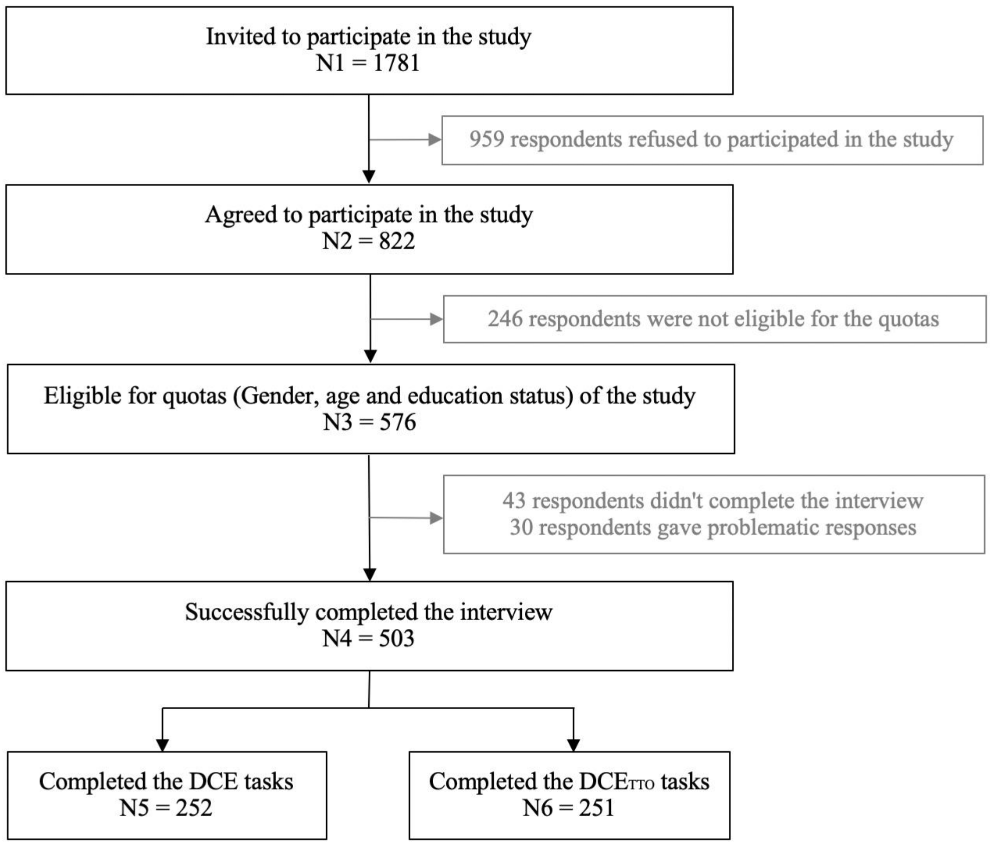

Fig.1 Flow chart of the sample inclusion

TTO, time trade-off; DCE, discrete choice experiment; DCETTO, discrete choice experiment with life duration.

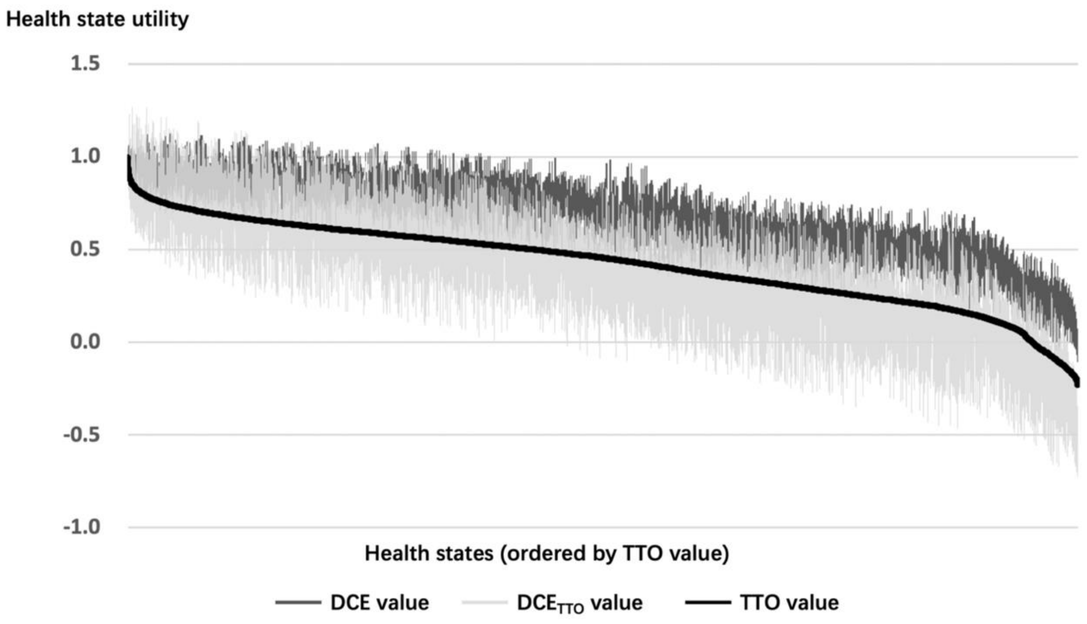

Fig. 2 A comparison among estimated values of 18,750 health states for three approaches

TTO, time trade-off; DCE, discrete choice experiment; DCETTO, discrete choice experiment with life duration.

Electronic Supplementary Material

Appendix Table 1 Comparison of characteristics between included and excluded respondents

Characteristics	Included respondents (N=503)	Excluded respondents (N=73)	P-value
Gender			
Male	270 (53.7%)	37 (50.7%)	0.613
Female	233 (46.3%)	36 (49.3%)	
Age (mean [SD])	45.4 (16.7)	49.2 (16.2)	**0.037**
Age group (y)			0.114
18-29	103 (20.5%)	7 (9.6%)	
30-39	100 (19.9%)	15 (20.5%)	
40-49	88 (17.5%)	14 (19.2%)	
50-59	94 (18.7%)	17 (23.3%)	
≥ 60	118 (23.4%)	20 (27.4%)	
Education			0.288
Primary or lower	93 (18.5%)	9 (12.3%)	
Junior high school	169 (33.6%)	31 (42.5%)	
Senior high school	115 (22.9%)	16 (21.9%)	
College or higher	126 (25.0%)	17 (23.3%)	
Ethnic group			**0.035**
Han Chinese	479 (95.2%)	73 (100.0%)	
Other	24 (4.8%)	0 (0%)	
Household registration			**0.029**
Urban	344 (68.4%)	58 (79.5%)	
Rural	159 (31.6%)	15 (20.5%)	
Marital status			0.615
Unmarried	111 (22.1%)	12 (16.4%)	
Married	352 (69.9%)	56 (76.7%)	
Divorced	15 (3.0%)	2 (2.7%)	
Widowed	25 (5.0%)	3 (4.1%)	
Health insurance			
Urban employee	312 (62.0%)	50 (68.5%)	0.242
Urban and rural resident	182 (36.2%)	18 (24.7%)	**0.034**
Commercial	93 (18.5%)	10 (13.7%)	0.155
Other	5 (1.0%)	2 (2.7%)	0.155
No	5 (1.0%)	2 (2.7%)	0.114
Employment status			0.451
Employed	297 (59.0%)	42 (57.5%)	
Retired	125 (24.9%)	23 (31.5%)	
Student	49 (9.7%)	5 (6.8%)	

Unemployed	32 (6.4%)	3 (4.1%)	
Monthly income (RMB)			0.946
< 2000	106 (21.0%)	14 (19.2%)	
2000-5000	293 (58.3%)	45 (61.6%)	
5000-10000	78 (15.5%)	11 (15.1%)	
>10000	26 (5.2%)	3 (4.1%)	
Smoking status			0.871
Never	331 (65.8%)	49 (67.1%)	
Former smoker	53 (10.5%)	6 (8.2%)	
Still	119 (23.7%)	18 (24.7%)	
Alcohol consumption			0.479
Never	277 (55.1%)	45 (61.6%)	
Former drinker	53 (10.5%)	6 (8.2%)	
Still	173 (34.4%)	22 (30.1%)	
Number of chronic conditions			0.638
0	294 (58.4%)	44 (60.3%)	
1	124 (24.7%)	17 (23.3%)	
2	44 (8.7%)	5 (6.8%)	
3	25 (5.0%)	4 (5.5%)	
4 or more	16 (3.2%)	3 (4.1%)	

[1] Among the 73 excluded respondents, 43 respondents were excluded because they did not complete the interview (9 for could not understand either of the three valuation tasks, 13 for interrupted by other persons, and 21 for did not have the patience to complete all the interview), and the other 30 respondents were excluded because they gave problematic responses (7 for gave all health states the same values in TTO tasks, 13 for less than 4 minutes in either of the three tasks, 4 for gave responses "AAAAA" or "BBBBB" in DCE tasks, and 6 for gave responses "AAAAA" or "BBBBB" in DCE_{TTO} tasks).

[2] The comparison of characteristics distributions between included and excluded respondents by t-test, chi2 or Fisher exact test as appropriate.

[3] The chronic conditions include: Hypertension, dyslipidemia, diabetes or high blood sugar, cancer or malignant tumor, chronic lung disease, liver disease, heart disease, stroke, kidney disease, stomach or other digestive disease, emotional or psychiatric problems, memory-related disease, arthritis or rheumatism, asthma, or other respondent-reported chronic conditions.

Appendix Table 2 Unadjusted estimated coefficients of the fitted models

	TTO data (N=503)		**DCE data (N=252)**		**DCE_{TTO} data [a] (N=251)**	
	Fixed effects model		**Conditional logit model (Latent utility)**		**Conditional logit model (Latent utility)**	
	Coef.	SE	Coef.	SE	Coef.	SE
Physical functioning						
PF2	-0.031	0.019	-0.171	0.107	-0.024	0.018
PF3	-0.039	0.023	-0.249*	0.102	-0.018	0.016
PF4	-0.135***	0.022	-0.406***	0.109	-0.090***	0.018
PF5	-0.411***	0.027	-1.796***	0.132	-0.169***	0.017
Role limitation						
RL2	-0.036	0.019	-0.037	0.106	**0.015**	0.017
RL3	-0.052*	0.023	-0.143	0.107	-0.020	0.018
RL4	-0.066**	0.020	-0.203	0.104	-0.039*	0.019
RL5	-0.088***	0.023	-0.533***	0.115	-0.044**	0.017
Social functioning						
SF2	-0.110***	0.021	**0.262****	0.089	**0.088*****	0.018
SF3	-0.112***	0.021	**0.340****	0.113	**-0.006**	0.018
SF4	-0.132***	0.020	-0.242*	0.109	**0.036***	0.015
SF5	-0.117***	0.020	-0.339**	0.109	**-0.023**	0.018
Pain						
PN2	-0.082***	0.023	**0.029**	0.102	-0.029	0.020
PN3	-0.088***	0.020	-0.161	0.110	-0.033	0.019
PN4	-0.076***	0.020	**0.062**	0.104	-0.060**	0.019
PN5	-0.334***	0.026	-1.315***	0.129	-0.167***	0.021
PN6	-0.351***	0.027	-1.691***	0.143	-0.199***	0.022
Mental health						
MH2	-0.037	0.019	-0.040	0.111	-0.048**	0.019
MH3	-0.117***	0.021	-0.218	0.113	-0.046**	0.018
MH4	-0.121***	0.023	-0.763***	0.116	-0.058***	0.017
MH5	-0.137***	0.022	-0.577***	0.116	-0.135***	0.020
Vitality						
VT2	-0.068***	0.020	**0.281***	0.114	-0.001	0.017
VT3	-0.061***	0.020	**0.106**	0.107	-0.033*	0.016
VT4	-0.114***	0.021	-0.220*	0.103	-0.086***	0.016
VT5	-0.125***	0.020	-0.433***	0.106	-0.093***	0.019
Life duration	--	--	--	--	0.384***	0.032
AIC	3209.645		4986.719		5221.096	
BIC	3373.446		5139.828		5490.647	

[a] In DCE_{TTO} data, the coefficients were for the interactions between dimension levels and life duration, for example, PF2*life duration.

The coefficients in bold meant non-monotonic with opposite sign; *** p <0.001, ** p <0.01, * p <0.05.

AIC, Akaike information criterion; BIC, Bayesian information criterion.

Appendix Table 3 Comparison of model coefficients between all data and data after exclusion [a]

	TTO data		DCE data		DCE_{TTO} data [b]	
	Before exclusion (N=533)	After exclusion (N=503)	Before exclusion (N=265)	After exclusion (N=252)	Before exclusion (N=268)	After exclusion (N=251)
Physical functioning						
PF2	-0.033	-0.031	-0.163	-0.171	-0.022	-0.024
PF3	-0.041	-0.039	-0.236	-0.249	-0.015	-0.018
PF4	-0.135	-0.135	-0.409	-0.406	-0.090	-0.090
PF5	-0.405	-0.411	-1.793	-1.796	-0.169	-0.169
Role limitation						
RL2	-0.037	-0.036	-0.020	-0.037	0.014	0.015
RL3	-0.054	-0.052	-0.133	-0.143	-0.021	-0.020
RL4	-0.071	-0.066	-0.213	-0.203	-0.043	-0.039
RL5	-0.094	-0.088	-0.519	-0.533	-0.045	-0.044
Social functioning						
SF2	-0.113	-0.110	0.252	0.262	0.084	0.088
SF3	-0.115	-0.112	0.325	0.340	-0.007	-0.006
SF4	-0.133	-0.132	-0.275	-0.242	0.034	0.036
SF5	-0.116	-0.117	-0.333	-0.339	-0.024	-0.023
Pain						
PN2	-0.083	-0.082	0.052	0.029	-0.030	-0.029
PN3	-0.090	-0.088	-0.131	-0.161	-0.033	-0.033
PN4	-0.074	-0.076	0.067	0.062	-0.056	-0.060
PN5	-0.337	-0.334	-1.307	-1.315	-0.167	-0.167
PN6	-0.352	-0.351	-1.682	-1.691	-0.197	-0.199
Mental health						
MH2	-0.041	-0.037	-0.018	-0.040	-0.048	-0.048
MH3	-0.119	-0.117	-0.194	-0.218	-0.048	-0.046
MH4	-0.122	-0.121	-0.726	-0.763	-0.056	-0.058
MH5	-0.139	-0.137	-0.553	-0.577	-0.133	-0.135
Vitality						
VT2	-0.067	-0.068	0.292	0.281	-0.002	-0.001
VT3	-0.060	-0.061	0.128	0.106	-0.035	-0.033
VT4	-0.107	-0.114	-0.207	-0.220	-0.087	-0.086
VT5	-0.124	-0.125	-0.417	-0.433	-0.092	-0.093
Life duration	--	--	--	--	0.387	0.384
AIC	3426.849	3209.645	5159.801	4986.719	5501.815	5221.096
BIC	3592.604	3373.446	5242.830	5139.828	5672.147	5490.647

[a] 30 respondents were excluded because they gave problematic responses (7 for gave all health states the same values in TTO tasks, 13 for less than 4 minutes in either of the three tasks, 4 for gave responses "AAAAA" or "BBBBB" in DCE tasks, and 6 for gave responses "AAAAA" or "BBBBB" in DCE_{TTO} tasks); [b] The coefficients for DCE_{TTO} data were for interactions between dimension levels and life duration, for example, PF2*life duration.AIC, Akaike information criterion; BIC, Bayesian information criterion.

Appendix Table 4 The acceptability of TTO, DCE and DCE$_{TTO}$ tasks in elderly (aged≥60 years) respondents

Characteristics Mean (SD) / N (%)	TTO tasks (N=118)	DCE tasks (N=57)	DCE$_{TTO}$ tasks (N=61)	p-value (TTO vs. DCE)	p-value (TTO vs. DCE$_{TTO}$)	p-value (DCE vs. DCE$_{TTO}$)
Completion time (min)	16.8 (8.6)	11.9 (5.4)	11.3 (4.7)	**<0.001**	**<0.001**	0.524
Self-reported difficulty level of understanding				0.985	0.229	0.270
Very easy	17 (14.4%)	8 (14.0%)	7 (11.5%)			
Easy	58 (49.2%)	27 (47.4%)	39 (63.9%)			
Moderate	33 (28.0%)	16 (28.1%)	9 (14.8%)			
Hard	9 (7.6%)	5 (8.8%)	6 (9.8%)			
Very hard	1 (0.8%)	1 (1.8%)	0 (0%)			
Self-reported difficulty level of answering				0.323	0.846	0.102
Very easy	11 (9.3%)	1 (1.8%)	9 (14.8%)			
Easy	62 (52.5%)	29 (50.9%)	32 (52.5%)			
Moderate	32 (27.1%)	17 (29.8%)	14 (23.0%)			
Hard	11 (9.3%)	8 (14.0%)	5 (8.2%)			
Very hard	2 (1.7%)	2 (3.5%)	1 (1.6%)			

Appendix Table 5 The acceptability of TTO, DCE and DCE$_{TTO}$ tasks in low education level (primary schools or lower) respondents

Characteristics Mean (SD) / N (%)	TTO tasks (N=93)	DCE tasks (N=46)	DCE$_{TTO}$ tasks (N=47)	p-value (TTO vs. DCE)	p-value (TTO vs. DCE$_{TTO}$)	p-value (DCE vs. DCE$_{TTO}$)
Completion time (min)	13.5 (6.7)	10.7 (5.6)	9.6 (5.0)	**0.015**	**<0.001**	0.331
Self-reported difficulty level of understanding				0.596	0.524	0.427
Very easy	5 (5.4%)	3 (6.5%)	2 (4.3%)			
Easy	41 (44.1%)	22 (47.8%)	26 (55.3%)			
Moderate	31 (33.3%)	12 (26.1%)	15 (31.9%)			
Hard	13 (14.0%)	9 (19.6%)	4 (8.5%)			
Very hard	3 (3.2%)	0 (0%)	0 (0%)			
Self-reported difficulty level of answering				0.370	0.278	0.441
Very easy	4 (4.3%)	3 (6.5%)	5 (10.6%)			
Easy	39 (41.9%)	20 (43.5%)	21 (44.7%)			
Moderate	33 (35.5%)	12 (26.1%)	16 (34.0%)			
Hard	11 (11.8%)	10 (21.7%)	5 (10.6%)			
Very hard	6 (6.5%)	1 (2.2%)	0 (0%)			

Appendix Table 6 The acceptability of TTO and DCE tasks in DCE group (N=252)

Characteristics	TTO tasks Mean (SD) / N (%)	DCE tasks Mean (SD) / N (%)	p-value
Completion time (min)	13.2 (7.7)	8.9 (4.5)	**<0.001**
Self-reported difficulty level of understanding			0.984
Very easy	27 (10.7%)	26 (10.3%)	
Easy	124 (49.2%)	127 (50.4%)	
Moderate	82 (32.5%)	79 (31.3%)	
Hard	15 (6.0%)	17 (6.7%)	
Very hard	4 (1.6%)	3 (1.3%)	
Self-reported difficulty level of answering			0.945
Very easy	21 (8.3%)	17 (6.7%)	
Easy	104 (41.3%)	107 (42.5%)	
Moderate	85 (33.7%)	87 (34.5%)	
Hard	33 (13.1%)	34 (13.5%)	
Very hard	9 (3.6%)	7 (2.8%)	

Appendix Table 7 The acceptability of TTO and DCE_{TTO} tasks in DCE_{TTO} group (N=251)

Characteristics	TTO tasks Mean (SD) / N (%)	DCE_{TTO} tasks Mean (SD) / N (%)	p-value
Completion time (min)	12.5 (6.4)	8.5 (5.6)	**<0.001**
Self-reported difficulty level of understanding			0.639
Very easy	36 (14.3%)	33 (13.1%)	
Easy	130 (51.8%)	139 (55.4%)	
Moderate	66 (26.3%)	63 (25.1%)	
Hard	17 (6.8%)	16 (6.4%)	
Very hard	2 (0.8%)	0 (0%)	
Self-reported difficulty level of answering			0.988
Very easy	34 (13.5%)	34 (13.5%)	
Easy	114 (45.4%)	115 (45.8%)	
Moderate	70 (27.9%)	73 (29.1%)	
Hard	26 (10.4%)	23 (9.2%)	
Very hard	7 (2.8%)	6 (2.4%)	

Appendix Table 8 Estimated model coefficients of TTO data for both DCE and DCE_{TTO} group

	TTO data of DCE group (N=252)			TTO data of DCE_{TTO} group(N=251)		
	Fixed effects model			Fixed effects model		
	Coef.	SE	p-value	Coef.	SE	p-value
Physical functioning						
PF2	0.037	0.033	0.254	0.032	0.032	0.325
PF3	0.032	0.034	0.509	0.047	0.035	0.107
PF4	0.122	0.030	<0.001	0.151	0.032	<0.001
PF5	0.406	0.031	<0.001	0.418	0.031	<0.001
Role limitation						
RL2	0.034	0.030	0.500	0.048	0.031	0.050
RL3	0.058	0.033	0.079	0.054	0.032	0.089
RL4	0.065	0.033	0.046	0.065	0.034	0.054
RL5	0.094	0.033	0.002	0.076	0.033	0.026
Social functioning						
SF2	0.094	0.029	0.004	0.119	0.030	<0.001
SF3	0.122	0.032	<0.001	0.104	0.031	0.002
SF4	0.110	0.031	<0.001	0.143	0.031	<0.001
SF5	0.113	0.031	<0.001	0.121	0.032	<0.001
Pain						
PN2	0.079	0.032	0.012	0.087	0.033	0.008
PN3	0.080	0.033	0.015	0.099	0.034	0.004
PN4	0.053	0.034	0.213	0.109	0.033	0.001
PN5	0.333	0.034	<0.001	0.331	0.036	<0.001
PN6	0.327	0.034	<0.001	0.380	0.033	<0.001
Mental health						
MH2	0.053	0.029	0.072	0.022	0.029	0.457
MH3	0.111	0.033	0.001	0.126	0.035	<0.001
MH4	0.124	0.032	<0.001	0.118	0.031	<0.001
MH5	0.147	0.032	<0.001	0.126	0.032	<0.001
Vitality						
VT2	0.067	0.030	0.028	0.075	0.032	0.017
VT3	0.068	0.033	0.037	0.056	0.033	0.087
VT4	0.111	0.031	<0.001	0.116	0.032	<0.001
VT5	0.136	0.033	<0.001	0.116	0.032	<0.001
AIC		1567.7220			1671.5950	
BIC		1679.9000			1778.0880	

Appendix Table 9 Comparison of estimated models for TTO data

	M1: OLS model			M2: FE model			M3: RE model		
	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
Intercept	-0.030	0.015	0.042	-0.024	0.020	0.238	-0.025	0.011	0.029
Physical functioning									
PF2	-0.047	0.021	0.029	-0.031	0.019	0.113	-0.034	0.018	0.056
PF3	-0.085	0.025	0.001	-0.039	0.023	0.094	-0.047	0.022	0.031
PF4	-0.147	0.026	<0.001	-0.135	0.022	<0.001	-0.137	0.021	<0.001
PF5	-0.449	0.034	<0.001	-0.411	0.027	<0.001	-0.417	0.028	<0.001
Role limitation									
RL2	-0.028	0.021	0.196	-0.036	0.019	0.054	-0.035	0.018	0.058
RL3	-0.064	0.024	0.008	-0.052	0.023	0.022	-0.055	0.022	0.012
RL4	-0.065	0.027	0.016	-0.066	0.020	0.001	-0.065	0.020	<0.001
RL5	-0.103	0.029	<0.001	-0.088	0.023	<0.001	-0.091	0.023	<0.001
Social functioning									
SF2	-0.107	0.023	<0.001	-0.110	0.021	<0.001	-0.109	0.020	<0.001
SF3	-0.108	0.027	<0.001	-0.112	0.021	<0.001	-0.110	0.021	<0.001
SF4	-0.131	0.027	<0.001	-0.132	0.020	<0.001	-0.131	0.021	<0.001
SF5	-0.091	0.023	<0.001	-0.117	0.020	<0.001	-0.113	0.020	<0.001
Pain									
PN2	-0.066	0.028	0.020	-0.082	0.023	<0.001	-0.079	0.023	0.001
PN3	-0.081	0.022	<0.001	-0.088	0.020	<0.001	-0.087	0.019	<0.001
PN4	-0.085	0.021	<0.001	-0.076	0.020	<0.001	-0.077	0.020	<0.001
PN5	-0.334	0.032	<0.001	-0.334	0.026	<0.001	-0.334	0.026	<0.001
PN6	-0.349	0.028	<0.001	-0.351	0.027	<0.001	-0.351	0.027	<0.001
Mental health									

MH2	-0.054	0.027	0.041	-0.037	0.019	0.058	-0.041	0.020	0.037
MH3	-0.060	0.027	0.027	-0.117	0.021	<0.001	-0.108	0.021	<0.001
MH4	-0.079	0.026	0.002	-0.121	0.023	<0.001	-0.114	0.022	<0.001
MH5	-0.143	0.031	<0.001	-0.137	0.022	<0.001	-0.139	0.023	<0.001
Vitality									
VT2	-0.047	0.022	0.034	-0.068	0.020	0.001	-0.064	0.018	<0.001
VT3	-0.061	0.025	0.015	-0.061	0.020	0.003	-0.062	0.020	0.002
VT4	-0.113	0.022	<0.001	-0.114	0.021	<0.001	-0.114	0.021	<0.001
VT5	-0.094	0.029	0.001	-0.125	0.020	<0.001	-0.119	0.020	<0.001
F-test		<0.001 (FE model were preferred)							
Hausman test						<0.001 (FE model were preferred)			
R^2		0.3270			0.3236			0.3246	
AIC		5988.0520			3209.6450			4730.9350	
BIC		6151.8530			3373.4460			4907.3360	
RMSD		0.1401			0.1465			0.1445	
MAD		0.0961			0.1008			0.0996	
ICC		0.9380			0.9380			0.9390	

Abbr: OLS model, ordinary least squares model; FE model, fixed-effect model; RE model, random-effect model. AIC, akaike information criterion; BIC, bayesian information criterion; RMSD, root mean squared difference; MAD, mean absolute difference; ICC, intraclass correlation coefficient.

Appendix Table 10 Comparison of estimated models for DCE data

	DCE data								
	Conditional logit model			**Mixed logit model**					
	Coef.	SE	p-value	Coef.	SE	p-value	SD	SE	p-value
Physical functioning									
PF2	-0.171	0.107	0.111	-0.148	0.115	0.201	0.314	0.295	0.287
PF3	-0.249	0.102	0.015	-0.077	0.111	0.486	0.064	0.351	0.854
PF4	-0.406	0.109	<0.001	-0.384	0.116	0.001	0.080	0.267	0.766
PF5	-1.796	0.132	<0.001	-1.213	0.140	<0.001	0.900	0.197	<0.001
Role limitation									
RL2	-0.037	0.106	0.726	0.336	0.119	0.005	--	--	--
RL3	-0.143	0.107	0.178	0.128	0.117	0.275	--	--	--
RL4	-0.203	0.104	0.051	-0.356	0.113	0.002	--	--	--
RL5	-0.533	0.115	<0.001	-0.173	0.111	0.119	--	--	--
Social functioning									
SF2	0.262	0.089	0.003	0.240	0.109	0.028	--	--	--
SF3	0.340	0.113	0.003	-0.107	0.112	0.340	--	--	--
SF4	-0.242	0.109	0.026	-0.032	0.110	0.767	--	--	--
SF5	-0.329	0.109	0.003	-0.395	0.108	<0.001	--	--	--
Pain									
PN2	0.029	0.102	0.772	0.049	0.128	0.917	0.774	0.220	0.917
PN3	-0.161	0.110	0.144	-0.044	0.121	0.716	0.171	0.209	0.412
PN4	0.062	0.104	0.553	-0.016	0.120	0.001	0.155	0.215	0.469
PN5	-1.315	0.129	<0.001	-1.075	0.127	<0.001	0.423	0.212	0.046
PN6	-1.691	0.143	<0.001	-1.482	0.154	<0.001	0.884	0.223	0.003
Mental health									
MH2	-0.040	0.111	0.719	-0.230	0.102	0.025	--	--	--
MH3	-0.218	0.113	0.054	-0.435	0.112	0.035	--	--	--
MH4	-0.763	0.116	<0.001	-0.397	0.110	<0.001	--	--	--
MH5	-0.577	0.116	<0.001	-0.764	0.124	<0.001	--	--	--
Vitality									
VT2	0.281	0.114	0.014	-0.058	0.114	0.164	0.183	0.200	0.360
VT3	0.106	0.107	0.323	0.075	0.117	0.018	0.102	0.230	0.659
VT4	-0.220	0.103	0.032	-0.386	0.113	0.001	0.449	0.180	0.067
VT5	-0.433	0.106	<0.001	-0.663	0.120	<0.001	0.702	0.168	<0.001
Log likelihood		-2463.3596				-2372.5185			
AIC		4986.719				4777.037			
BIC		5139.828				4964.585			

The coefficients of all levels in one dimension was set as random coefficients if the estimated standard deviation of any one level in this dimension was statistically significant ($p<0.05$). This study tested a number of sets of the coefficients, and the model which set all of the levels in RL, SF and MH as fixed coefficients and set the rest of the levels as random coefficients, was selected as the best model in terms of Log likelihood, AIC and BIC.

The conditional logit model was selected as the better model in terms of the less non-monotonic coefficients. Besides, not large heterogeneity based on a few coefficients with statistically significant SD were found in mixed logit model.

Appendix Table 11 Comparison of estimated models for DCE_{TTO} data

	DCE_{TTO} data								
	Conditional logit model			**Mixed logit model**					
	Coef.	SE	p-value	Coef.	SE	p-value	SD	SE	p-value
Year	0.384	0.031	<0.001	0.565	0.054	0.000	0.267	0.026	<0.001
Physical functioning*Year									
PF2	-0.024	0.018	0.167	-0.026	0.026	0.327	0.083	0.044	0.058
PF3	-0.018	0.016	0.247	0.012	0.025	0.646	0.001	0.052	0.988
PF4	-0.090	0.018	<0.001	-0.093	0.027	0.001	0.129	0.030	0.064
PF5	-0.169	0.017	<0.001	-0.276	0.030	<0.001	0.165	0.034	<0.001
Role limitation *Year									
RL2	0.015	0.017	0.373	0.006	0.026	0.804	--	--	--
RL3	-0.020	0.018	0.246	-0.025	0.027	0.343	--	--	--
RL4	-0.039	0.019	0.037	-0.071	0.026	0.007	--	--	--
RL5	-0.044	0.017	0.008	-0.059	0.025	0.021	--	--	--
Social functioning*Year									
SF2	0.088	0.018	<0.001	0.050	0.027	0.064	0.106	0.041	0.089
SF3	-0.006	0.018	0.752	-0.004	0.026	0.879	0.112	0.035	0.048
SF4	0.036	0.015	0.019	-0.003	0.026	0.879	0.048	0.040	0.228
SF5	-0.023	0.018	0.207	-0.059	0.025	0.016	0.023	0.040	0.571
Pain*Year									
PN2	-0.029	0.020	0.156	-0.039	0.030	0.190	0.114	0.040	0.004
PN3	-0.033	0.019	0.081	-0.028	0.030	0.362	0.072	0.046	0.121
PN4	-0.060	0.019	0.002	-0.048	0.027	0.077	0.022	0.045	0.633
PN5	-0.167	0.021	<0.001	-0.240	0.030	<0.001	0.102	0.033	0.287
PN6	-0.199	0.022	<0.001	-0.319	0.036	<0.001	0.235	0.047	0.095
Mental health*Year									
MH2	-0.048	0.019	0.012	-0.059	0.026	0.060	--	--	--
MH3	-0.046	0.018	0.009	-0.050	0.025	0.045	--	--	--
MH4	-0.058	0.017	0.001	-0.107	0.025	<0.001	--	--	--
MH5	-0.135	0.020	<0.001	-0.205	0.029	<0.001	--	--	--
Vitality*Year									
VT2	-0.001	0.017	0.933	-0.028	0.026	0.274	--	--	--
VT3	-0.033	0.016	0.037	-0.079	0.027	0.011	--	--	--
VT4	-0.086	0.016	<0.001	-0.078	0.025	0.002	--	--	--
VT5	-0.093	0.019	<0.001	-0.122	0.027	<0.001	--	--	--
Log likelihood	-2634.5479			-2410.292					
AIC	5221.0960			4932.584					
BIC	5490.6470			5212.558					

The coefficients of all levels in one dimension was set as random coefficients if the estimated standard deviation of any one level in this dimension was statistically significant ($p<0.05$). This study tested a number of sets of the coefficients, and the model which set all of the levels in RL, MH and VT as fixed coefficients and set the rest of the levels as random coefficients, was selected as the best model in terms of Log likelihood, AIC and BIC.

The conditional logit model was selected as the better model in terms of the less non-monotonic coefficients. Besides, not large heterogeneity based on a few coefficients with statistically significant SD were found in mixed logit model.

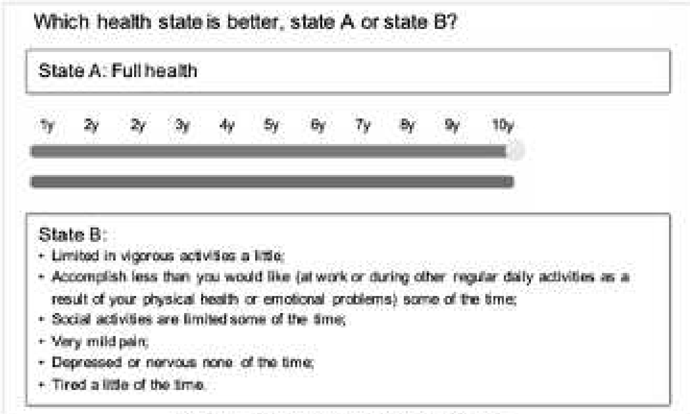

Fig 1a The example of the TTO task

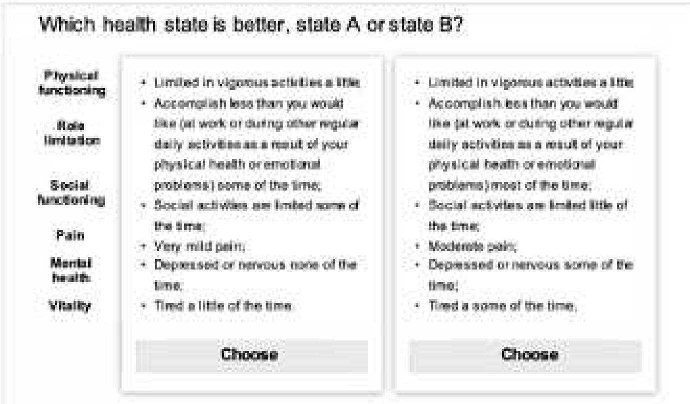

Fig 1b The example of the DCE task

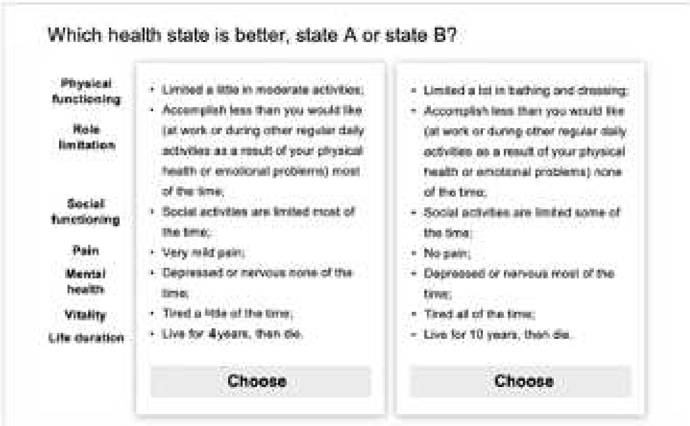

Fig 1c The example of the DCE_{TTO} task

Appendix Fig. 1 The examples of the translated elicitation tasks used in the study