Quality of Life Research (2022) 31:2791–2803
https://doi.org/10.1007/s11136-022-03159-2

# Discrete choice experiment with duration versus time trade-off: a comparison of test–retest reliability of health utility elicitation approaches in SF-6Dv2 valuation

**Shitong Xie[1,2] · Jing Wu[2,3] · Gang Chen[4]**

Accepted: 5 May 2022 / Published online: 25 May 2022


## Abstract

**Objectives** To evaluate and compare the test–retest reliability of discrete choice experiments with duration ($DCE_{TTO}$) and time trade-off (TTO) in the Chinese SF-6Dv2 valuation study.

**Methods** During face-to-face interviews, a representative sample of the Chinese general population completed 8 TTO tasks and 10 $DCE_{TTO}$ tasks. Retest interviews were conducted after two weeks. For both $DCE_{TTO}$ and TTO, the consistency of raw responses between the two tests was firstly evaluated at the individual level. Regressions were conducted to investigate the association between the test–retest reliability and the respondents' characteristics and the severity of health states. Consistency was then analyzed at the aggregate level by comparing the rank order of the coefficients of dimensions.

**Results** In total, 162 respondents (51.9% male; range 18–80 years) completed the two tests. The intraclass correlations coefficient 0.958 for TTO, with identical values accounting for 59.3% of observations. 76.4% of choices were identical for $DCE_{TTO}$, with a Kappa statistic of 0.528. Respondents' characteristics had no significant impact while the severity of health states valued in TTO and $DCE_{TTO}$ tasks had a significant impact on the test–retest reliability. Both approaches produced relatively stable rank order of dimensions in constrained model estimations between test and retest data.

**Conclusions** Individual responses of both approaches are relatively stable over time. The rank orders of dimensions in model estimations between test and retest for TTO and $DCE_{TTO}$ are also consistent. The differences of utility estimation between the two tests for $DCE_{TTO}$ need to be further investigated based on a larger sample size.

**Keywords** Test–retest reliability · Health state valuation · Discrete choice experiment · Time trade-off · SF-6D · China

✉ Jing Wu
jingwu@tju.edu.cn

✉ Gang Chen
gang.chen@monash.edu

1 Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

2 School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China

3 Center for Social Science Survey and Data, Tianjin University, Tianjin, China

4 Monash Business School, Centre for Health Economics, Monash University, Melbourne, VIC, Australia

## Introduction

The quality-adjusted life year (QALY) is regarded as one of the most important outcomes in economic evaluations of healthcare interventions [1]. It is calculated by multiplying a quality adjustment weight (or health utility) by life duration to generate a standardized metric that can then be used in cost-utility analysis (CUA) [1]. A common approach to eliciting the health utility values is the use of generic preference-based measures, such as the EQ-5D or SF-6D [1–3]. A generic preference-based measure usually consists of a health state classification system and a corresponding country-specific health utility value set elicited from a representative sample of the general population [1].

The health state utility values have been widely elicited using cardinal approaches, such as standard gamble (SG) and time trade-off (TTO) [1, 4, 5]. However, 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 old adults or children [6]. One of the most recent developments in utility elicitation is the adoption of the discrete choice experiment (DCE), especially for online surveys [7–10]. The DCE with duration ($DCE_{TTO}$) approach, a variant of DCE, provides a novel alternative to elicit the utility values [10, 11]. Unlike traditional DCE, in which only different hypothetical health states are presented, $DCE_{TTO}$ requires respondents to further consider the duration of living in each hypothetical health state, i.e., it includes quantity vs. quality trade-off in each task. Consequently, it does not require a separate task to anchor the latent utility, which remains controversial in the traditional DCE approach [6, 12]. Compared with the iterative process of identifying the indifference point between two options in SG and TTO, $DCE_{TTO}$ is usually regarded as a promising alternative since it only requires respondents to make ordinal choices [10].

The evidence on reliability is crucial when assessing the performance of an elicitation approach [1, 13]. Assessment of reliability commonly refers to two different types of validation [14]. The first one is called internal reliability which mainly assesses the homogeneity of multi-item scales and is not the focus of this study. The second one, test–retest reliability, focuses on the repeatability and stability of measurements. Test–retest reliability is a critical property to ensure that the elicited societal preference on different health states is stable over time. So far, the evidence on test–retest reliability of different elicitation approaches, including visual analogue scale (VAS), SG, TTO, DCE, and $DCE_{TTO}$, is very limited [13, 15–20]. Overall, mixed results were found in the studies that compared VAS, TTO, and SG, based on the reported intraclass correlation coefficients (ICC) [16–19]. The two studies that compared traditional DCE and TTO also reported mixed results on test–retest reliability [13, 20]. Currently, no studies compared the test–retest reliability of $DCE_{TTO}$ with other approaches. A summary of existing evidence on test–retest reliability of different elicitation approaches can be found in Table 1.

Given the increasing usage of the $DCE_{TTO}$ in health state valuation [21], it is crucial to deepen our understanding of its test–retest reliability, in particular when comparing to the traditional approaches such as TTO. This study aimed to evaluate and compare the test–retest reliability of $DCE_{TTO}$ and TTO based on the SF-6Dv2 valuation tasks among a representative sample of the Chinese general population.

## Methods

This study was part of a larger study that focused on the valuation of the SF-6Dv2 using face-to-face interviews among the Chinese general population [22]. More detailed descriptions of the design of the valuation study can be found elsewhere [22].

### Instrument

The SF-6D is derived from the Short-Form 36 (SF-36) health survey [23]. The original health state classification system of the SF-6D comprises six dimensions with four to six levels in each, including physical functioning (PF), role limitation (RL), social functioning (SF), pain (PN), mental health (MH), and vitality (VT), yielding up to 18,000 health states [23]. Recently, a second version of the SF-6D, SF-6Dv2, was developed, which revisited the items selected from the SF-36 and modified the ambiguity between dimension levels and inconsistency of wording in the original version [24]. The SF-6Dv2 has the same six dimensions with five to six levels in each dimension, resulting in 18,750 health states in total [24–26]. The Simplified Chinese version of the SF-6Dv2 was developed after translation and cross-cultural adaption, and preliminary psychometric testing was conducted among the Chinese general population [26].

### Elicitation tasks design

The composite TTO approach (hereafter TTO) [10, 22, 27–29] and $DCE_{TTO}$ elicitation approaches were employed in this study (Supplementary Fig. 1) [22]. A total of 295 states were selected for TTO tasks, including the six mildest imperfect states, the worst state, and 288 other states generated based on near orthogonal arrays using SAS® Studio. These 288 states were firstly divided into 48 blocks, the worst state (included in all 48 blocks) and the six mildest states (each randomly included in eight blocks) were then added in these blocks. Respondents were randomly assigned to 1 of the 48 blocks for valuation.

For $DCE_{TTO}$ tasks, four levels of life duration, i.e., 1, 4, 7, and 10 years, were chosen [22, 25]. The $DCE_{TTO}$ tasks, which consisted of 300 state pairs distributed over 30 blocks, were generated using the balanced overlap method, with the maximized statistical efficiency according to the D-efficiency based on Lighthouse Studio 9.6.0 (Sawtooth Software, Inc) [22, 30–32]. The task order and the left–right position of health states within each task were all randomized. Respondents were randomly assigned to 1 of the 30 blocks for valuation.

### Sample and data collection

Respondents included in this study were recruited from eight cities and their surrounding rural areas to cover a wide geographical range with varying economic development stages in China [22]. A quota sampling method was used to recruit a representative sample of the Chinese general population

**Table 1** A summary of the comparison of the test–retest reliability of elicitation approaches in previous studies

| Authors | Year | Population | Sample size | Administration mode | Description of health states | Elicitation approaches | Interval between two tests (days) | Reported ICC | Conclusion |
|---|---|---|---|---|---|---|---|---|---|
| Purba et al. [13] | 2018 | Representative sample of the Indonesian general population | 226 | Face-to-face | EQ-5D-5L | TTO, DCE | Mean:19.9; SD: 9.32; Range: 10–59 | TTO: -0.16–0.81; DCE: NA | TTO is stable over time, whereas in DCE the relative values of the dimensions shift |
| Gamper et al. [15] | 2018 | Representative sample of the German and French general population | German: 300; French: 305 | Online | QLU-C10D | $DCE_{TTO}$ | Range: 28–42 | Germany: 0.790–0.796; France: 0.857–0.879 | The individual choices are sufficiently stable over time to support the validity of $DCE_{TTO}$ |
| Kim et al. [18] | 2017 | Sample of the Korean general population | 105 | Face-to-face | EQ-5D-5L | VAS, SG, TTO | 14 | VAS: 0.906; SG: 0.841; TTO: 0.827 | The test–retest reliability of the three methods was acceptable for eliciting the preference scores |
| Robinson et al. [17] | 2011 | Random sample of population lived in Birmingham, UK | TTO: 171; PTO: 151 | Postal | EQ-5D-3L | TTO, PTO | Range: 14–28; (60% at 14) | TTO: 0.40–0.88; PTO: -0.17–0.82 | While the reliability results for TTO were generally positive, the reliability results for PTO are less clear |
| Bijlenga et al. [20] | 2009 | Sample of community population in Netherlands | 97 | Face-to-face | Vignettes on moderate-risk pregnancy at term | VAS, TTO, DCE | Median: 5; Range: 3–21 | VAS: 0.77; TTO: 0.79; DCE: NA | The test–retest reliability was high and comparable throughout the three methods, but interobserver consistency was different: high for VAS, slightly lower for DCE, and low for TTO |
| Lin MR, et al. [19] | 2006 | Patients with traumatic spinal cord injuries in Taiwan, China | 20 | Telephone | Current state | VAS, SG, TTO | 14 | VAS: 0.89–0.92; SG: 0.73–0.78; TTO: 0.78–0.91 | The VAS generally performed better than the SG and TTO among people with traumatic spinal cord injuries |

**Table 1** (continued)

| Authors | Year | Population | Sample size | Administration mode | Description of health states | Elicitation approaches | Interval between two tests (days) | Reported ICC | Conclusion |
|---|---|---|---|---|---|---|---|---|---|
| Badia et al. [16] | 1999 | Random sample of the Spanish population | 50 | Face-to-face | EQ-5D-3L | VAS, TTO | 7–28 | VAS: 0.90 (95%CI 0.88–0.92); TTO: 0.84 (95%CI 0.81–0.87) | Both methods demonstrated high test–retest reliability, with the VAS providing slightly higher reliability than the TTO |

*ICC* intraclass correlation coefficients, *TTO* time trade-off, *DCE* discrete choice experiment, *$DCE_{TTO}$* discrete choice experiment with duration, *VAS* visual analogue scale, *PTO* person trade-off, *SG* standard gamble

in terms of age, gender, and area of residence (urban/rural) [33, 34]. Face-to-face, computer-based interviews were conducted in this study.[1] More detailed information can be found in the main valuation paper [22].

After the first interview (*test*), the interviewers asked for the respondents' consent to re-participate in the face-to-face interview again (*retest*) and collected their contact information. While the interval between test and retest was set as two weeks [15, 35], it could be relaxed to the range of 10–30 days to ensure that respondents could be interviewed again at their convenient time. The retest interview was held using the same process by the same interviewers. In the retest interview, respondents were assigned to the same block of tasks for both TTO and $DCE_{TTO}$, with the same previously described randomization of the order of tasks, as in the first interview.

## Data analysis

Descriptive analyses were first conducted to present the respondents' characteristics, as well as the distributions of both TTO and $DCE_{TTO}$ data. The utility values of the respondents' self-reported SF-6Dv2 health states were calculated using the Chinese-specific value set [22]. Then, the test–retest data for two approaches were analyzed at both the individual level and aggregate level.

For the $DCE_{TTO}$ tasks, a calculation of the "pseudo-QALY" approach[2] was employed to present the relative preference for choice A versus B in each choice pair. The pseudo-QALY was obtained by multiplying the utility value of the health state (calculated using the Chinese-specific $DCE_{TTO}$ value set of SF-6Dv2 [22]) by the corresponding life duration. For example, the difference of pseudo-QALY for choice A (121122 with 4 years) and B (413334 with 1 year) in a $DCE_{TTO}$ task would be (0.971 * 4)–(0.639 * 1) = 3.245 pseudo-QALYs.

### Statistical analyses at the individual level

For the TTO data, three evaluations were conducted. First, the number of respondents changing 0–8 out of their 8 responses between test and retest was investigated. Second, the proportion of responses that had different values

[1] The process of the first interview was as follows [21]: respondents (1) completed inclusion and quota questions, to confirm s/he was eligible; (2) reported their health using the SF-6Dv2; (3) completed the TTO and $DCE_{TTO}$ tasks with the order randomized; and (4) reported a series of social-demographic characteristics.

[2] For traditional DCE tasks, the distribution of relative preference for choice A versus B could be observed by evaluating the difference in the severity of the health states (i.e., the severity score of the health state) included in both choices [13]. However, this approach is not applicable in this study given there exists additional life duration dimension in the $DCE_{TTO}$ task.

between the two tests was evaluated. Any significant difference in values between test and retest was assessed applying the Wilcoxon matched-pairs signed-rank test. The degree of consistency between two tests was also evaluated using the intraclass correlations coefficient (ICC), with the interpretation of ICC < 0.40 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, and > 0.74 = excellent [36]. Third, the degree of agreement of TTO observed values between test and retest was assessed by the Bland–Altman plot.

For the $DCE_{TTO}$ data, first, the number of respondents changing 0–10 out of their 10 choices between test and retest was investigated. Second, the proportion of choices that were identical between the two tests was evaluated. The overall agreement irrespective of respondents and blocks was calculated, with the good agreement confirmed at ≥ 70% [37]. The kappa ($\kappa$) statistic was also calculated to provide the estimation of agreement that is corrected for chance, with the interpretation ($\kappa$) < 0.40 = low, 0.41–0.60 = moderate, 0.61–0.80 = good, and > 0.80 = excellent [15, 38]. Third, the proportions of respondents that gave consistent choices between two tests in different pseudo-QALYs categories were shown using a histogram.

The performance of test–retest reliability among subgroups with different demographic characteristics, the time intervals and the difference in self-reported utility value between two tests was also evaluated. Linear regression was used for TTO data, with the dependent variable being the difference in observed TTO values between the two tests. A binary logistic regression model was used for $DCE_{TTO}$ data, in which the dependent variable was measured by whether or not identical choices were observed between the two tests. Cluster-robust standard errors were used to account for one respondent completing multiple tasks.

#### Statistical analyses at the aggregate level

Considering the relatively small sample size of this study, constrained main-effect only model specifications were estimated for TTO and $DCE_{TTO}$ data, respectively. Different from the main valuation study, in which a set of dummy variables was used for each dimension [22], here each dimension was modeled as a continuous variable.

Equation (1) was used to model TTO data:

$$y_i = \alpha + \beta_1 PF + \beta_2 RL + \beta_3 SF + \beta_4 PN + \beta_5 MH + \beta_6 VT + \epsilon \quad (1)$$

where $y_i$ is the disutility value given by the respondent $i$; α is the intercept; *PF*, *RL*, *SF*, *PN*, *MH* and *VT* are continuous variables representing the different levels in each dimension of SF-6Dv2, assuming linear effect across levels; $\beta$ are the estimated coefficients on each dimension; and $\varepsilon$ is the error term.

Equation (2) was used to model $DCE_{TTO}$ data:

$$U_{ij} = \lambda_0 t_{ij} + \lambda_1 PFt_{ij} + \lambda_2 RLt_{ij} + \lambda_3 SFt_{ij} + \lambda_4 PNt_{ij} + \lambda_5 MHt_{ij} + \lambda_6 VTt_{ij} + \varepsilon_{ij} \quad (2)$$

where $U_{ij}$ is the binary choice of respondent $i$ for $DCE_{TTO}$ task $j$; $t_{ij}$ is the life duration, which is modeled as a linear variable; $\lambda_0$ is the coefficient for the life duration; *PF*, *RL*, *SF*, *PN*, *MH* and *VT* are continuous variables representing the different levels in each dimension, assuming linear effect across levels (and they were included as interaction terms with the life duration variable); correspondingly $\lambda$ are coefficients for the interactions; $\varepsilon_{ij}$ is the error term.

Both test and retest data for TTO and $DCE_{TTO}$ were modeled using the optimal statistical methods that were selected to generate the Chinese-specific value set of SF-6Dv2 [22]. In brief, the TTO data were analyzed using the random-effect model; the $DCE_{TTO}$ data were analyzed using a conditional logit model, following the model specification proposed by Bansback et al. [10] and the corresponding method of anchoring on the QALY scale [10, 39–42]. More detailed information can be found in the main valuation paper [22]. Owing to the sample size of the retest data, the consistency between test and retest was mainly focused on the rank order of SF-6Dv2 dimensions from the model estimations. The scatter plot was also drawn to visually demonstrate the degree of consistency.

All statistical analyses were conducted using STATA 15.1. For the comparison of distributions of characteristics between subgroups, the t-test was used for continuous variables, while the $\chi^2$ or Fisher exact test was used for categorical variables. A two-tailed $p$-value < 0.05 was considered statistically significant.

## Results

### Respondents

Of 178 respondents who consented to participate in the retest survey, 16 respondents were excluded because they did not complete the second interview. Consequently, 162 respondents were included in this study. The mean (standard deviation, SD) interval between the first and the second interviews was 15.6 (4.4) days (range 10–33 days). As illustrated in Table 2, the mean (SD) age of the sample was 44.4 (16.5) years, ranging from 18 to 80 years, 51.9% were males, and 37.7% lived in rural areas. The distributions of characteristics of the respondents were similar to those of the Chinese general population in terms of age, gender, and proportion of urban/rural residence [33, 34]. The utility values of self-reported health state using SF-6Dv2 in both interviews were 0.868 and 0.872, respectively. The absolute mean (SD) difference of utility value between the two interviews was 0.026 (0.052), with a range of 0–0.428.

## TTO data

A total of 1,296 TTO responses were provided by the 162 respondents for each test. Histograms of the TTO observed values showed a comparable distribution between both tests (Fig. 1). More than half of the respondents ($N=118$, 72.8%) changed four or less of their eight responses in the retest, with only five (3.1%) respondents changing more than seven responses (Supplementary Table 1). Of the 1,296 responses, 770 (59.4%) were identical between the two tests, 231 (17.8%) increased, and 295 (22.8%) decreased (Supplementary Table 3). While the mean absolute difference between the two tests ranged from 0 to 0.142 with an average mean (SD) absolute difference of 0.029 (0.081), there was only one health state (555655) with a significant change ($p=0.041$) in median value (Supplementary Table 3). The ICC ranged from 0.500 to 1.000, with a mean ICC of 0.945. The Bland–Altman plot (Fig. 2) showed that the mean difference of observed TTO values between test and retest was 0.01. The 95% limits of agreement ranged from −0.17 to 0.19, and 92.2% of points lay within limits.

## $DCE_{TTO}$ data

The $DCE_{TTO}$ data consisted of 1,620 responses per test. Histograms of the relative preference for choice A vs. B by the difference in pseudo-QALYs for the two tests showed a similar expected distribution (Fig. 1), in which respondents were always more likely to choose the choice that had a more pseudo-QALYs. 116 (71.6%) respondents gave three or fewer different responses among 10 tasks (Supplementary Table 2). Only two (1.2%) respondents gave seven different responses, and no respondents gave eight or more different responses. The overall agreement was 76.4%, with 1,238 of 1,620 responses being identical between the two tests (Supplementary Table 3). The kappa ($\kappa$) statistic was 0.528, which was interpreted as a moderate agreement [15, 38]. As shown in Fig. 3, the proportions of respondents who gave consistent choices between two tests in different pseudo-QALYs categories ranged from 66.7 to 86.4%. A slightly higher proportion could be observed among categories with larger differences of pseudo-QALYs between the two choices.

## Subgroup analyses

As illustrated in Table 3, the differences in observed TTO values between the two tests were not statistically significant

**Table 2** The Characteristics of respondents

| Characteristics | Chinese general population[a] (%) | Total sample (N = 162) N (%) |
|---|---|---|
| Gender | | |
| Male | 51.2% | 84 (51.9%) |
| Female | 48.8% | 78 (48.1%) |
| Age (mean [SD]) | NA | 44.4 (16.5) |
| Age group (y) | | |
| 18–29 | 21.5% | 43 (26.5%) |
| 30–39 | 18.7% | 16 (9.9%) |
| 40–49 | 21.1% | 31 (19.1%) |
| 50–59 | 17.1% | 28 (17.3%) |
| ≥ 60 | 21.6% | 44 (27.2%) |
| Education | | |
| Primary school or lower | 26.2% | 56 (34.6%) |
| Middle school | 40.3% | 28 (17.3%) |
| High school | 17.2% | 33 (20.4%) |
| College or higher | 16.3% | 45 (27.8%) |
| Region | | |
| Urban | 59.6% | 101 (62.3%) |
| Rural | 40.4% | 61 (37.7%) |
| Marital status | | |
| Unmarried | 18.6% | 46 (28.4%) |
| Married | 73.9% | 109 (67.3%) |
| Divorced | 2.0% | 3 (1.9%) |
| Widowed | 5.5% | 4 (2.5%) |
| Health insurance | | |
| Urban employee | NA | 88 (54.3%) |
| Urban and rural resident | NA | 65 (40.1%) |
| Commercial | NA | 22 (13.6%) |
| Other | NA | 3 (1.9%) |
| No | NA | 3 (1.9%) |
| Employment status | | |
| Employed | NA | 96 (59.3%) |
| Retired | NA | 28 (17.3%) |
| Student | NA | 25 (15.4%) |
| Unemployed | NA | 13 (8.0%) |
| Monthly income (RMB) | | |
| < 2000 | NA | 48 (29.6%) |
| 2000–5000 | NA | 83 (51.2%) |
| 5000–10,000 | NA | 21 (13.0%) |
| > 10,000 | NA | 10 (6.2%) |
| Number of chronic conditions[b] | | |
| 0 | NA | 96 (59.3%) |
| 1 | NA | 40 (24.7%) |
| 2 | NA | 12 (7.4%) |
| 3 | NA | 8 (4.9%) |
| ≥ 4 | NA | 6 (3.7%) |
| Self-reported SF-6Dv2 utility value[c] (mean [SD]) | | |
| Test | NA | 0.868 (0.115) |
| Retest | NA | 0.872 (0.109) |

**Table 2** (continued)

[a]Statistics data of Chinese general population were extracted from the Sixth National Census of China (2010), the China Statistical Yearbook 2018, and the Statistical bulletin on national economic and social development of China (2018). When the statistical scale of the original data was not calculated as the general population aged ≥ 18 years, the data were adjusted based on the proportion of the population of each age to the total population in this study. N/A indicates that data was not included in the public available data source

[b]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 diseases, emotional or psychiatric problems, memory-related disease, arthritis or rheumatism, asthma, or other respondent-reported chronic conditions

[c]The utility value was calculated using the Chinese SF-6Dv2 value set [21]. The result of the t-test of the self-reported SF-6Dv2 utility values between the two tests was not statistically significant ($p = 0.375$)

among subgroups of all characteristics, except for the difference in the self-reported SF-6Dv2 utility values between the two tests ($p < 0.001$). The differences in the proportion of identical choices for $DCE_{TTO}$ were statistically significant only between subgroups with or without chronic conditions ($p = 0.029$). Linear regression analysis demonstrated that the difference in observed TTO value between the two tests became larger when the severity score of the health state valued in the TTO task increased (Coef. = 0.001, 95% CI: [0.000, 0.002], $p$-value = 0.025) (Supplementary Table 4). Logistic regression for the $DCE_{TTO}$ data showed that respondents were more likely to give consistent choices between the two tests in the task with a larger difference of pseudo-QALYs (Coef. = 0.112, 95% CI: [0.048, 0.176], $p$-value = 0.001) (Supplementary Table 4).

### Comparisons on aggregated model estimates

As shown in Table 4, the constrained models for TTO data showed a consistent rank order of dimensions between test and retest, i.e., PN > PF > MH > RL > VT > SF, with all coefficients being statistically significant. Similarly, the constrained models for $DCE_{TTO}$ data also showed a consistent rank order of PN > PF > MH > SF > VT > RL, while the coefficients of RL and VT were not statistically significant in both test and retest models (Table 4). The scatter plots (Supplementary Fig. 2) demonstrated that, while generally good consistency was observed for both approaches, the consistency of the estimated utility values between the two tests for TTO was slightly higher than that for $DCE_{TTO}$.

## Discussion

When compared with the most widely used approach TTO, $DCE_{TTO}$ is commonly regarded as a promising alternative [10]. To the best of our knowledge, this study provided the first empirical evidence that directly compared the test–retest reliability between TTO and $DCE_{TTO}$ approaches. The results demonstrated good test–retest reliability of both utility elicitation approaches in the context of developing the SF-6Dv2 value set in China. Moreover, it should be borne in mind that the implications of the observed levels of test–retest reliability for the two approaches are different, since TTO values are directly modeled as utility values, while $DCE_{TTO}$ is modeled as latent values under random utility theory.

The test–retest reliability reported in this study is comparable to or better than what has been reported in the literature. The ICC for TTO was 0.945, which is higher than the previous studies that ranged from 0.780 to 0.880 [16–20]; 59.4% of the responses were identical between the two tests, which is also higher than 24.5% reported in a previous study [13]. Regarding the $DCE_{TTO}$, the overall agreement (76.4 vs. 70.6–80.2%) and the kappa ($\kappa$) statistic (0.528 vs. 0.411–0.605) are consistent or higher than those from the previous study [15]. The better result on both TTO and $DCE_{TTO}$ data may be partly due to the different interview methods employed in these studies, i.e., the face-to-face interview in this study and postal or telephone interviews in most of the previous studies [15, 17, 19]. It is also worth noting that more respondents reported extreme values of both the worst (−1) and the best TTO values (1) in the retest than those reported in the test, with lower mean values (0.356 vs. 0.367) in the retest. This finding was not consistent with a previous study, which reported a higher mean value of 0.042 in the retest [13]. With very limited studies focusing on test–retest reliability of valuation techniques, more studies are warranted to confirm this finding.

Findings from subgroups analyses in this study are worth highlighting. In regression analyses for both TTO and $DCE_{TTO}$ data, characteristics of respondents, including age, gender, education level, chronic condition status, marital status, regions of residence, and the self-reported utility values had no statistically significant impact on the TTO observed values or the $DCE_{TTO}$ choices between the two tests. There was a significant negative effect of the severity score on the health state valued in the TTO task. The finding is not surprising since the cognitive burden may be heavier when the health state in TTO tasks becomes worse, especially for those health states that are considered worse than

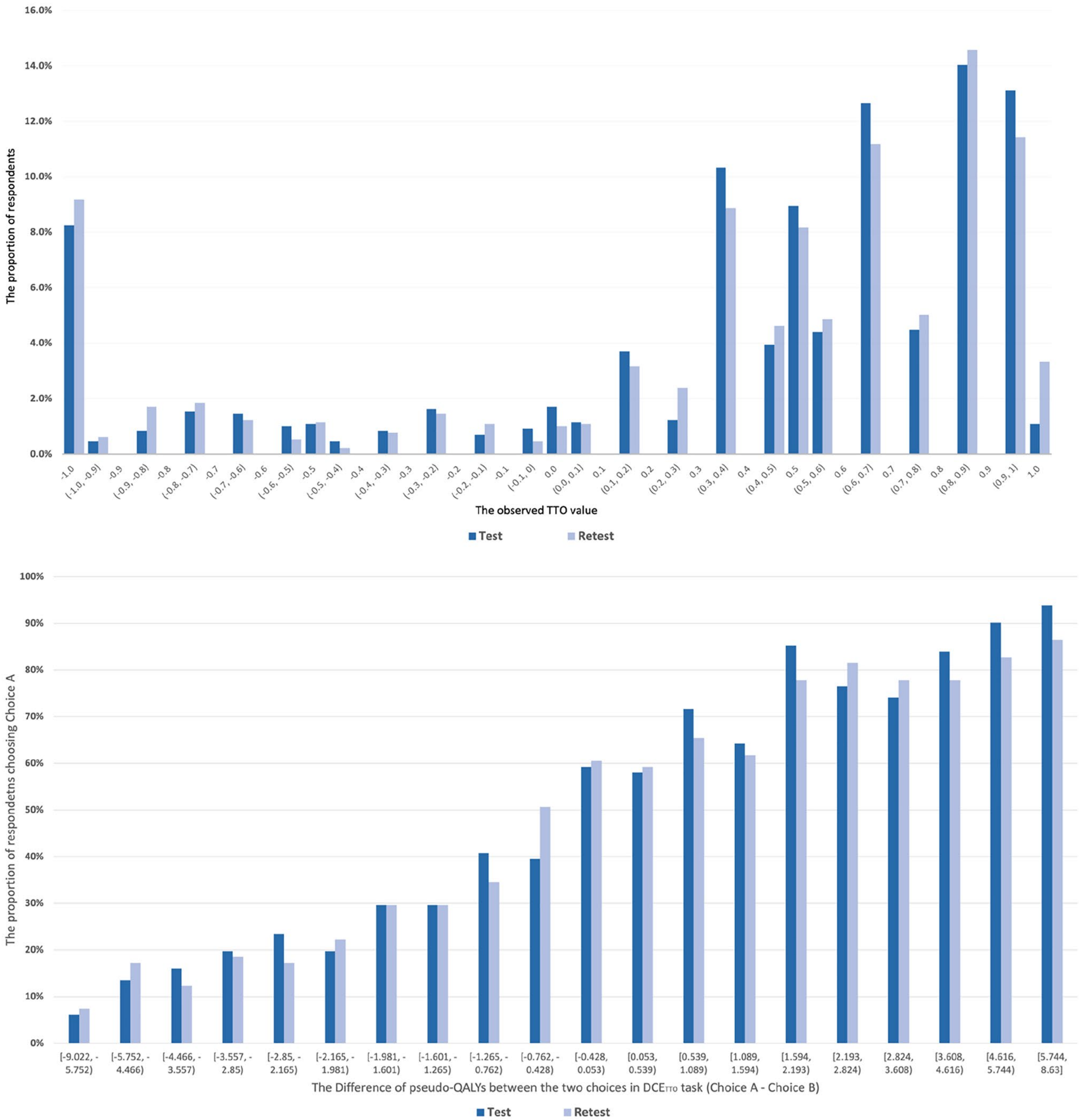
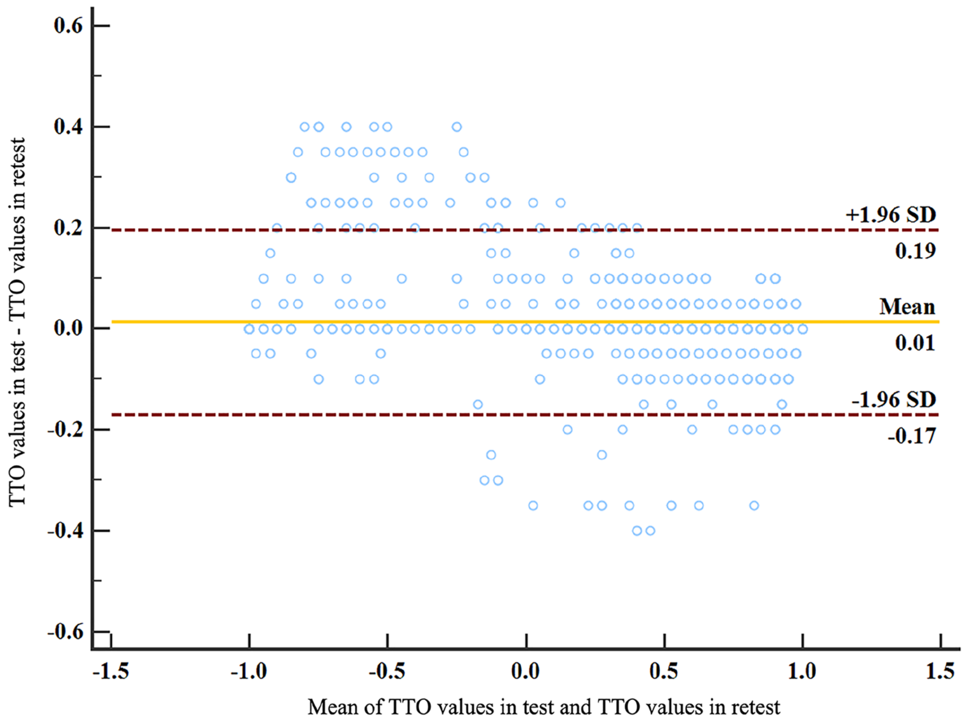



**Fig. 1** The comparison of response distribution for both test and retest data. The pseudo-QALYs was calculated by multiplying the utility value of the health state by the corresponding life duration. The utility value was calculated using the Chinese $DCE_{TTO}$ value set [21]. For example, the pseudo-QALY for choice A (121122 with 4 years) in a $DCE_{TTO}$ task would be 0.971 * 4 = 3.884 QALYs. *TTO* time trade-off, $DCE_{TTO}$ discrete choice experiment with duration

death, in which a different question design from states better than death is used [17, 28]. The finding for $DCE_{TTO}$ that it is easier to make a consistent choice when facing a larger difference of pseudo-QALYs between two options also seems potentially reasonable. These findings provide the first empirical exploration to evaluate the relationship between the social-demographic characteristics of respondents and the characteristics of health states and the test–retest reliability of utility elicitation approaches.

The time interval between test and retest was commonly considered associated with the reliability of the results. In this study, we found that the time interval between the

**Fig. 2** The Bland–Altman plot of TTO observed values for test and retest



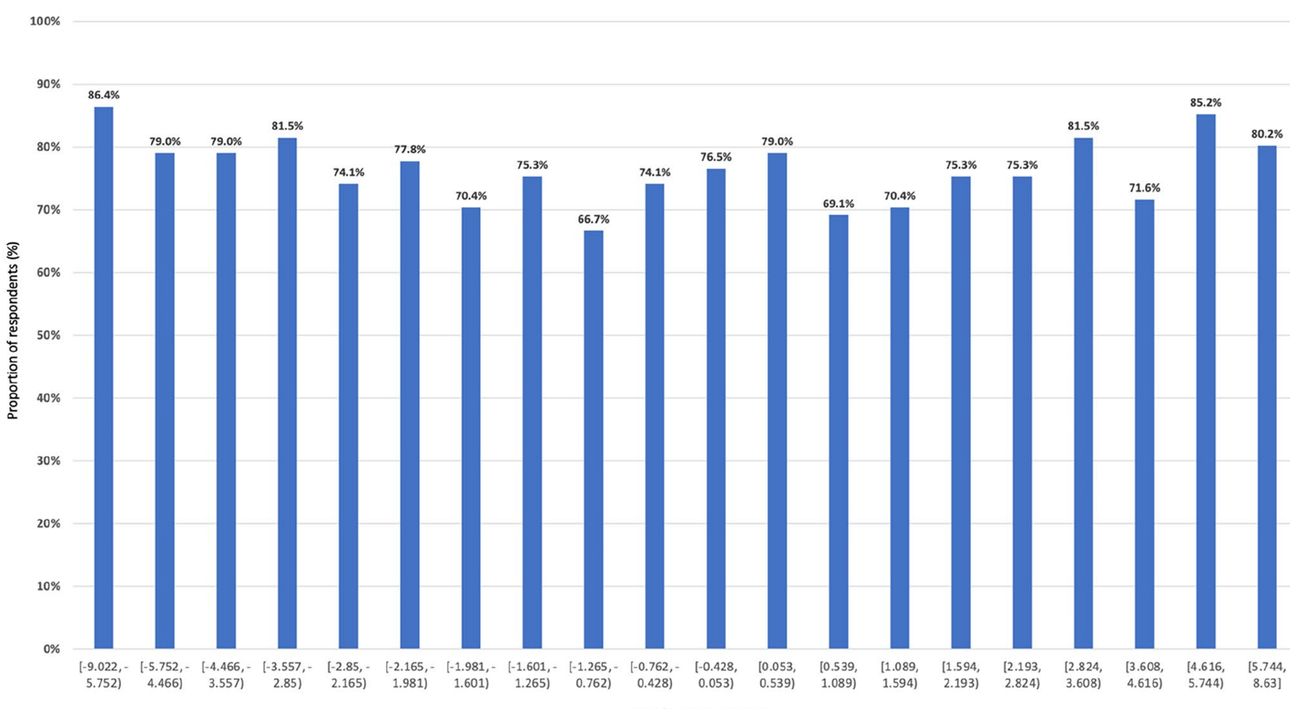




**Fig. 3** The proportions of respondents gave consistent choices in different pseudo-QALYs categories for $DCE_{TTO}$ data. The pseudo-QALYs was calculated by multiplying the utility value of the health state by the corresponding life duration. The utility value was calculated using the Chinese $DCE_{TTO}$ value set [21]. For example, the pseudo-QALY for choice A (121122 with 4 years) in a $DCE_{TTO}$ task would be 0.971 * 4 = 3.884 QALYs

**Table 3** Subgroup analysis for TTO and $DCE_{TTO}$ data

| Respondent | TTO data | | | | | $DCE_{TTO}$ data | |
|---|---|---|---|---|---|---|---|
| | N (%) | Mean (SD) | | Mean difference (SD) | ***p*-value**[b] | Identical between test and retest (N [%]) | ***p*-value**[c] |
| | | Test | Retest | | | | |
| Age group (50) | | | | | | | |
| Age < 50 | 72 (44.4%) | 0.415 (0.554) | 0.407 (0.575) | 0.007 (0.085) | 0.084 | 535 (74.3%) | 0.073 |
| Age ≥ 50 | 90 (55.6%) | 0.308 (0.632) | 0.291 (0.663) | 0.017 (0.101) | | 703 (78.1%) | |
| Age group (60) | | | | | | | |
| Age < 60 | 118 (72.8%) | 0.397 (0.567) | 0.386 (0.593) | 0.012 (0.090) | 0.93 | 911 (77.2%) | 0.224 |
| Age ≥ 60 | 44 (27.2%) | 0.288 (0.649) | 0.275 (0.676) | 0.012 (0.101) | | 327 (74.3%) | |
| Gender | | | | | | | |
| Male | 84 (51.9%) | 0.382 (0.568) | 0.370 (0.593) | 0.012 (0.091) | 0.972 | 645 (76.8%) | 0.719 |
| Female | 78 (48.1%) | 0.352 (0.618) | 0.340 (0.644) | 0.012 (0.095) | | 593 (76.0%) | |
| Education | | | | | | | |
| Below high school | 84 (51.9%) | 0.325 (0.609) | 0.312 (0.633) | 0.013 (0.095) | 0.637 | 635 (75.6%) | 0.417 |
| High school or above | 78 (48.1%) | 0.412 (0571) | 0.402 (0.598) | 0.011 (0.091) | | 603 (77.3%) | |
| Region | | | | | | | |
| Urban | 101 (62.3%) | 0.375 (0.600) | 0.365 (0.625) | 0.009 (0.091) | 0.275 | 782 (77.4%) | 0.22 |
| Rural | 61 (37.7%) | 0.356 (0.578) | 0.340 (0.606) | 0.015 (0.097) | | 456 (74.8%) | |
| Chronic conditions[a] | | | | | | | |
| None | 96 (59.3%) | 0.415 (0.539) | 0.405 (0.566) | 0.010 (0.091) | 0.586 | 752 (78.3%) | **0.029** |
| Any | 66 (40.7%) | 0.297 (0.656) | 0.283 (0.681) | 0.014 (0.096) | | 486 (73.6%) | |
| Marital status | | | | | | | |
| Married | 109 (67.3%) | 0.350 (0.610) | 0.335 (0.637) | 0.015 (0.095) | 0.123 | 831 (76.2%) | 0.805 |
| Other | 53 (32.7%) | 0.404 (0.552) | 0.398 (0.574) | 0.006 (0.089) | | 407 (76.8%) | |
| Time interval between two tests | | | | | | | |
| ≤ 14 days | 77 (47.5%) | 0.362 (0.578) | 0.345 (0.609) | 0.016 (0.094) | 0.125 | 585 (76.0%) | 0.744 |
| 15–19 days | 58 (35.8%) | 0.330 (0.633) | 0.321 (0.654) | 0.009 (0.096) | | 444 (76.6%) | |
| ≥ 20 days | 27 (16.7%) | 0.472 (0.509) | 0.467 (0.533) | 0.005 (0.085) | | 212 (78.5%) | |
| Difference of self-reported utility value between two tests | | | | | | | |
| 0 | 59 (36.4%) | 0.333 (0.618) | 0.329 (0.642) | 0.004 (0.092) | **< 0.001** | 451 (76.4%) | 0.983 |
| (0–0.026) | 55 (34.0%) | 0.392 (0.575) | 0.376 (0.602) | 0.016 (0.085) | | 419 (76.2%) | |
| > 0.026 | 48 (29.6%) | 0.382 (0.578) | 0.365 (0.607) | 0.017 (0.102) | | 368 (76.7%) | |

[a]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 diseases, emotional or psychiatric problems, memory-related disease, arthritis or rheumatism, asthma, or other respondent-reported chronic conditions

[b]The mean difference of TTO observed values between subgroups was tested by the Student's t test or $Chi^2$ tests as appropriate

[c]The difference of the distribution for $DCE_{TTO}$ data between subgroups was tested using $chi^2$ test

*TTO* time trade-off, $DCE_{TTO}$ discrete choice experiment with duration

two tests also had no significant impact on both TTO and $DCE_{TTO}$ approaches. Note that the mean time interval of 15.6 days in this study was comparable with previous studies, which mainly ranged from 3 to 59 days, with the means of 5–19 days [13, 16–20]. Further studies with larger time intervals are warranted to evaluate the relationship between time interval and consistency, as well as the memorizing effect (i.e., the respondents may remember their choices made for the elicitation tasks during the first test for a few days) on the test–retest reliability of the health utility elicitation approaches.

Both TTO and $DCE_{TTO}$ are relatively stable overtime on the rank order of dimensions in model estimations between test and retest, which provides evidence of feasibility in eliciting utility at the aggregate level for both approaches. We mainly focused on the constrained models, since the aim of this study was to compare the reliability of these approaches over time rather than generate the utility value set. Constrained models with fewer parameters could generate more

**Table 4** Estimated coefficients of the constrained models on TTO and $DCE_{TTO}$ data

| | TTO data | | | | | $DCE_{TTO}$ data | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Test | | Retest | | | Test | | Retest | |
| | Coef | SE | Coef | SE | | Coef | SE | Coef | SE |
| Intercept | 0.222*** | 0.029 | 0.231*** | 0.032 | Year | 0.621*** | 0.043 | 0.546*** | 0.043 |
| Physical functioning | −0.061*** | 0.011 | −0.059*** | 0.011 | Physical functioning*Year | −0.037*** | 0.005 | −0.041*** | 0.005 |
| Role limitation | −0.034*** | 0.007 | −0.037*** | 0.008 | Role limitation*Year | 0.001 | 0.005 | 0.005 | 0.005 |
| Social functioning | −0.024** | 0.007 | −0.022** | 0.008 | Social functioning*Year | −0.012* | 0.005 | −0.008* | 0.004 |
| Pain | −0.077*** | 0.008 | −0.079*** | 0.008 | Pain*Year | −0.054*** | 0.005 | −0.055*** | 0.005 |
| Mental health | −0.038*** | 0.009 | −0.042*** | 0.010 | Mental health*Year | −0.021*** | 0.005 | −0.009* | 0.005 |
| Vitality | −0.033*** | 0.008 | −0.036*** | 0.008 | Vitality*Year | −0.008 | 0.005 | −0.006 | 0.004 |

**p*-value < 0.05; ***p*-value < 0.01; ****p*-value < 0.001

robust results comparing a full model with four or five dummies for each dimension among a relatively small sample size. It is also worth noting that potentially better performance of the consistency of estimated values between test and retest data for TTO than that for $DCE_{TTO}$ were observed according to the scatter plots. While due to the relatively small sample size in this study, further study with a larger sample size is warranted to further evaluate the consistency of model estimations.

Several limitations of this study should be considered. First, considering the relatively small sample size and corresponding small number of observations for each task in this study given the same experimental designs as the Chinese SF-6Dv2 valuation study, there could be an impact on the statistical efficiency of the utility model estimation. The constrained model specifications (each dimension was modelled as a continuous variable) were therefore used in this study instead of using a full main-effects model (a set of dummy variables was used for each dimension). Second, the distribution of education level of the sample in this study was different from that of the Chinese general population, i.e., a higher proportion of respondents with college or higher degrees and primary school or lower education, and a lower proportion of junior high school were found in this study. However, the subgroup analysis for the education level demonstrated that this difference had a trivial impact on the study findings. Third, 16 interviewers employed in this study had the same extensive training but came from different regions of China, had different academic backgrounds, and they might adopt different interview skills, all of which might influence the findings that have been reported. However, the interview effect was negligible in this study by checking the distributions of TTO and $DCE_{TTO}$ data among different interviewers and cities. Fourth, there might be a selection effect for the respondents who completed the retest interviews. However, there was no significant difference in most demographic characteristics between the respondents included in the larger valuation study and respondents included in this study, except for the education level and employment status (Supplementary Table 5) [22]. Fifth, it should be noted that the computation of the pseudo-QALY approach might be distorted by time preferences. Subsequent studies using this approach should pay attention to this issue.

## Conclusions

Individual responses to both TTO and $DCE_{TTO}$ approaches are relatively stable over time. The rank orders of dimensions in model estimations between test and retest for TTO and $DCE_{TTO}$ are also consistent, which provides evidence of feasibility in eliciting utility at the aggregate level for both approaches. Subgroup analyses from this study demonstrated the potentially negligible relationship between the demographic characteristics of respondents and the test–retest reliability of both approaches. The differences in utility estimation between the two tests for $DCE_{TTO}$ need to be further investigated based on larger sample size.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11136-022-03159-2.

**Acknowledgements** This study was funded by the National Natural Science Foundation of China (grant No. 71673197 and No. 72174142). We would like to thank all the interviewers and respondents for taking part in this study.

**Author contributions** Concept and design: SX, JW, GC. Acquisition of data: SX, JW. Analysis and interpretation of data: SX, JW, GC. Drafting of the manuscript: SX, GC. Statistical analysis: SX, GC. Obtaining funding: JW. Supervision: JW. All authors commented on previous versions of the manuscript and approved the final manuscript.

## Declarations

**Conflict of interest** JW reported receiving grants from the National Natural Science Foundation of China during the conduct of the study. No other conflicts of interest were reported by the authors.

**Ethical approval** This study was approved by the Institutional Review Board of School of Pharmaceutical Science and Technology, Tianjin University (No. 20180615) and was conducted in accordance with the Declaration of Helsinki.

**Consent to participate** Informed consent was obtained from all individual participants included in the study. Participants were informed about their freedom of refusal. Anonymity and confidentiality were maintained throughout the research process.

**Consent to publish** Not applicable.

## References


1. Brazier, J., Ratcliffe, J., Saloman, J., & Tsuchiya, A. (2017). *Measuring and valuing health benefits for economic evaluation*. Oxford University Press.
2. Group, T. E. (1990). EuroQol-a new facility for the measurement of health-related quality of life. *Health Policy, 16*(3), 199–208.
3. Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference-based single index from the UK SF-36 health survey. *Journal of Clinical Epidemiology, 51*(11), 1115–1128.
4. Martin, A. J., Glasziou, P. P., Simes, R. J., & Lumley, T. (2000). A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. *International Journal of Technology Assessment in Health Care, 16*(1), 137–147.
5. Morimoto, T., & Fukui, T. (2002). Utilities measured by rating scale, time trade-off, and standard gamble: Review and reference for health care professionals. *Journal of Epidemiology, 12*(2), 160–178.
6. Brazier, J., Rowen, D., Yang, Y., & Tsuchiya, A. (2012). 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, 13*(5), 575–587.
7. Lancsar, E., & Louviere, J. (2008). Conducting discrete choice experiments to inform healthcare decision making: A user's guide. *PharmacoEconomics, 26*(8), 661–677.
8. Craig, B. M., & Busschbach, J. J. (2009). The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. *Population Health Metrics, 7*, 3.
9. Stolk, E. A., Oppe, M., Scalone, L., & Krabbe, P. F. M. (2010). Discrete choice modeling for the quantification of health states: The case of the EQ-5D. *Value in Health, 13*(8), 1005–1013.
10. Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics, 31*(1), 306–318.
11. Mulhern, B., Bansback, N., Hole, A. R., & Tsuchiya, A. (2017). Using discrete choice experiments with duration to model EQ-5D-5L health state preferences: Testing experimental design strategies. *Medical Decision Making, 37*(3), 285–297.
12. Rowen, D., Brazier, J., & Van Hout, B. (2015). A comparison of methods for converting DCE values onto the full health-dead QALY scale. *Medical Decision Making, 35*(3), 328–340.
13. Purba, F. D., Hunfeld, J. A. M., Timman, R., Iskandarsyah, A., Fitriana, T. S., Sadarjoen, S. S., et al. (2018). Test-retest reliability of EQ-5D-5L valuation techniques: The composite time trade-off and discrete choice experiments. *Value in Health, 21*(10), 1243–1249.
14. Fayers, P. M., & Machin, D. (2015). *Quality of life: The assessment, analysis and reporting of patient-reported outcomes* (3rd ed.). John Wiley & Sons.
15. Gamper, E.-M., Holzner, B., King, M. T., Norman, R., Viney, R., Nerich, V., & Kemmler, G. (2018). Test-retest reliability of discrete choice experiment for valuations of QLU-C10D health states. *Value in Health, 21*(8), 958–966.
16. Badia, X., Monserrat, S., Roset, M., & Herdman, M. (1999). Feasibility, validity and test-retest reliability of scaling methods for health states: The visual analogue scale and the time trade-off. *Quality of Life Research, 8*(4), 303–310.
17. Robinson, S. (2011). Test-retest reliability of health state valuation techniques: The time trade off and person trade off. *Health Economics, 20*(11), 1379–1391.
18. Kim, S. H., Lee, S. I., & Jo, M. W. (2017). Feasibility, comparability, and reliability of the standard gamble compared with the rating scale and time trade-off techniques in Korean population. *Quality of Life Research, 26*(12), 3387–3397.
19. Lin, M. R., Hwang, H. F., Chung, K. P., Huang, C., & Chen, C. Y. (2006). Rating scale, standard gamble, and time trade-off for people with traumatic spinal cord injuries. *Physical Therapy, 86*(3), 337–344.
20. Bijlenga, D., Birnie, E., & Bonsel, G. J. (2009). Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome vignettes on moderate-risk pregnancy at term. *Value in Health, 12*(5), 821–827.
21. Mulhern, B., Norman, R., Street, D. J., & Viney, R. (2019). One method, many methodological choices: A structured review of discrete-choice experiments for health state valuation. *PharmacoEconomics, 37*(1), 29–43.
22. Wu, J., Xie, S., He, X., Chen, G., Bai, G., Feng, D., et al. (2021). Valuation of SF-6Dv2 health states in China using time trade-off and discrete-choice experiment with a duration dimension. *PharmacoEconomics, 39*(5), 521–535.
23. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics, 21*(2), 271–292.
24. Brazier, J. E., Mulhern, B. J., Bjorner, J. B., Gandek, B., Rowen, D., Alonso, J., Vilagut, G., & Ware, J. E. (2020). Developing a new version of the SF-6D health state classification system from the SF-36v2: SF-6Dv2. *Medical Care, 58*(6), 557–565.
25. Mulhern, B. J., Bansback, N., Norman, R., & Brazier, J. (2020). Valuing the SF-6Dv2 classification system in the United Kingdom using a discrete-choice experiment with duration. *Medical Care, 58*(6), 566–573.
26. Wu, J., Xie, S., He, X., Chen, G., & Brazier, J. E. (2020). The Simplified Chinese version of SF-6Dv2: Translation, cross-cultural adaptation and preliminary psychometric testing. *Quality of Life Research, 29*(5), 1385–1391.
27. Janssen, B. M., Oppe, M., Versteegh, M. M., & Stolk, E. A. (2013). Introducing the composite time trade-off: A test of feasibility and face validity. *The European Journal of Health Economics, 14*(Suppl 1), S5-13.
28. Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goni, J. M., & Luo, N. (2016). EuroQol protocols for time trade-off valuation of health outcomes. *PharmacoEconomics, 34*(10), 993–1004.
29. Xie, S., Wu, J., He, X., Chen, G., & Brazier, J. E. (2020). Do discrete choice experiments approaches perform better than time trade-off in eliciting health state utilities? Evidence from SF6Dv2 in China. *Value in Health, 23*(10), 1391–1399.
30. Chrzan, K., & Orme, B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Sawtooth software research paper series*, 98382.

31. Johnson, F. R., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., Bresnahan, B. W., Kanninen, B., & Bridges, J. F. (2013). Constructing experimental designs for discrete-choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in Health, 16*(1), 3–13.
32. Marshall, D. A., Deal, K., Bombard, Y., Leighl, N., MacDonald, K. V., & Trudeau, M. (2016). 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, 6*(6), e010981.
33. National Bureau of Statistics of China. (2011). *China Sixth National Census 2010*. Retrieved April 4, 2022, from: http://www.stats.gov.cn/ztjc/zdtjgz/zgrkpc/dlcrkpc/.
34. National Bureau of Statistics of China. (2019). *China Statistical Yearbook 2018*. Retrieved April 4, 2022, from: http://www.stats.gov.cn/tjsj/ndsj/2018/indexeh.htm.
35. Schmelkin Pedhazur, L. (1991). *Measurement, design, and analysis: An integrated approach*. Psychology Press.
36. Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.
37. Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10*(1), 141–150.
38. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20*(1), 37–46.
39. Norman, R., Viney, R., Brazier, J., Burgess, L., Cronin, P., King, M., Ratcliffe, J., & Street, D. (2014). Valuing SF-6D health states using a discrete choice experiment. *Medical Decision Making, 34*(6), 773–786.
40. Viney, R., Norman, R., Brazier, J., Cronin, P., King, M. T., Ratcliffe, J., & Street, D. (2014). An Australian discrete choice experiment to value eq-5d health states. *Health Economics, 23*(6), 729–742.
41. King, M. T., Viney, R., Simon Pickard, A., Rowen, D., Aaronson, N. K., Brazier, J. E., et al. (2018). 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, 36*(2), 225–238.
42. Rowen, D., Mulhern, B., Stevens, K., & Vermaire, J. H. (2018). Estimating a Dutch value set for the pediatric preference-based CHU9D using a discrete choice experiment with duration. *Value in Health, 21*(10), 1234–1242.