CALY-SWE value set: An integrated approach for a valuation study based on an online-administered TTO and DCE survey

Authors:

Kaspar Walter Meili (Corresponding author),
Department of Epidemiology and Global Health
Umeå university, Sweden
kaspar.meili@umu.se, kaspar.meili@yahoo.de
Orcid: 0000-0002-9889-4406

Brendan Mulhern,
Centre for Health Economics Research and Evaluation
University of Technology Sidney, Australia
Orcid: 0000-0003-3656-8063

Richard Ssegonja,
Department of Public Health and Caring Sciences
Uppsala University, Sweden
Department of Medical Sciences,
Respiratory, Allergy and Sleep medicine research unit,
Uppsala University, Sweden
Orcid: 0000-0002-5323-5626

Jan Hjelte,
Department of Social work
Umeå university, Sweden
Orcid: 0000-0002-5269-1961

Kerstin Edin
Department of Epidemiology and Global Health
Umeå university, Sweden

Fredrik Norström,
Department of Epidemiology and Global Health
Umeå university, Sweden
Orcid: 0000-0002-0457-2175

Inna Feldman,
Department of Public Health and Caring Sciences
Uppsala University, Sweden
Department of Epidemiology and Global Health

Umeå university, Sweden
Orcid: 0000-0003-3329-6066

Anna Månsdotter,
Department of Epidemiology and Global Health
Umeå university, Sweden

Lars Lindholm
Department of Epidemiology and Global Health
Umeå university, Sweden
Orcid: 0000-0002-1633-2179

Abstract

Purpose: To explore and develop methods to use TTO and DCE data collected in self-administered online surveys to elicit a CALY-SWE value set.

Methods: Building on existing methodological knowledge around DCE and TTO studies, we optimized the web survey in an integrated approach that consisted of a qualitative face validity study, iterative web survey development including a three-stage roll-out, a customized experimental design, a sample size simulation. Based on the inconsistencies of TTO answers per participant, we assessed TTO data quality by calculating a score, and examined the effect of excluding TTO data according to this score on the value set modelling.

Results: Participants in the quality study informed improvements in the survey's visual design and phrasing. Based on the sample size simulation, we judged a sample size of 1500 with a balance of six DCE and five TTO questions to be appropriate for the valuation study. Change made for the second stage, for example the introduction of a *learning* state and of color-coding, improved TTO data quality. Excluding TTO answers per-participant based on the score lead to an improved TTO data foundation for the model by several metrics such as no inconsistent coefficients and reduced standard errors relative to the coefficient magnitude.

Conclusion: Developing value sets was feasible with online administered DCE and TTO questions if the web survey is sufficiently optimized and coherent with the experimental design. The severity of inconsistencies could be used to identify and exclude poor quality TTO data to strengthen the value set modelling.

Plain English summary

CALY-SWE is a new instrument for measuring quality of life broadly and for use in economic cost-effectiveness evaluations of social welfare interventions. To be used as such, a vale set is needed to assign quality weights to the states of live used in the economic evaluations. However, no methodology exists yet for CALY-SWE to derive a value set, and existing similar valuation studies from the health context use costly person-to-person interviews.

In this study, we aimed to find appropriate methods for deriving a CALY-SWE value set using an online survey.

We developed an online web survey with discrete choice experiment (DCE) and time trade-off (TTO) questions. We used previous methodological knowledge from similar studies, qualitative interview, a statistical simulation for the sample size, and we rolled-out the survey in three stages to be able to implement further improvements. We also sought a way to identify and exclude participants who contributed poorer quality TTO answers to improve the underlying data for the value set.

The resulting survey was optimised for online administration with a shorter survey length of six DCE and five TTO questions, compensated by an increased sample size of 1500. We developed a score based on illogical TTO answers that allows to identify and exclude poorer quality TTO data. This design made it possible to perform the CALY-SWE valuation study.

Keywords: quality-adjusted life year, time trade-off, discrete choice experiment, capability approach, online survey, economic evaluation

Word count: 4000 = 4123 – 46 (f3) – 54 (f2) – 23 (f1)

Introduction

CALY-SWE (capability-adjusted life years Sweden) is a new measure for quality of life purposed for use in economic evaluations with broad social consequences [1, 2]. CALY-SWE conceptually relates to the quality-adjusted life year (QALY) concept developed within health economics [1, 2], and a value set is needed to use the instrument in economic cost-effectiveness evaluations [3]. Value sets consists of quality scores, or weights, on the [0,1] scale for all *states* that the instrument describes. CALY-SWE consists of 6 attributes (health, social relations, financial situation and housing, occupation, political and civil rights, and security) where each has 3 levels (Do not agree, Partially agree, Agree completely) , equalling 729 possible states [2]. Zero corresponds to a quality of life equivalent to death, and 1 to a quality of life sufficient for a flourishing life [4].

Traditionally in health economics, standard gamble and time trade-off (TTO) questions have been widely used for measuring health state values. More recently, discrete choice experiment (DCE) questions have also been widely adopted in the development of value sets [3].

Both TTO and DCE questions provide complementary information on preferences. In DCE questions, participants ordinally compare two states with each other, providing information on the relative strength of levels and attributes. Results are, however, only anchored relative to each other and not on the absolute [0,1] scale. TTO questions evaluate a single state by comparing time spent in the best state with time in an impaired state. The time in the best state is then gradually changes until the participant states that both situations are equivalent. The quality weight of the impaired state equals the time spent in the impaired state divided by the time spent in the best state. TTO questions thus yield direct information on the absolute anchoring of states on the [0,1] scale but are cognitively challenging [5, 6]. Recently, hybrid models [7, 8] have been developed that can jointly estimate value sets from DCE and TTO data, making it possibly to integrate the complementary preference information of DCE and TTO.

EQ-5D is a widely used preference-based measure where country specific value are derived with the EQ-VT protocol [9] for person-to-person interviews. The EQ-VT protocol has been refined over time and provides routines for both DCE and TTO questions, defines their experimental design, the interview procedures and in parallel provides software for doing the interviews. The protocol also focuses on interview training and data quality monitoring per interviewer to increase data quality, for example it defines a minimum time to spend in the introductory TTO question. However, the EQ-VT is designed for person-to person interviews and unsupervised self-administered surveys are not intended. Person-to-person interviews increase the costs and requirements for sampling and data collection considerably compared to traditional self-administered surveys, thus decreasing the overall feasibility. This especially applies in the context of COVID-19 restrictions which happened to coincide with the CALY-SWE valuation study.

On the other hand, some evidence suggests that face-to-face supervised administration leads to better data quality compared to unsupervised online administration for TTO questions [10]. Data quality issues related to the challenging TTO format may also be exacerbated by using a commercial panel of participants together with online self-administration.

For the CALY-SWE valuation study we decided to rely on an online panel and a self-administrated survey with TTO and DCE questions as we judged it to be the most feasible way to time- and cost- effectively collect data representative of the Swedish population. However, given the evidence on issues related to online self-administration, we implemented and tested multiple elements to adapt the methods and increase their validity.

Aim

Our aim was to explore and develop methods for eliciting a CALY-SWE value set, based on online data collection via a web panel, and by leveraging methodological experience accumulated from existing studies using both TTO and DCE.

To overcome the challenging nature of data quality in online surveys, especially for TTO questions, we adopted an integrated approach over several areas, which translated into the specific goals of:

1) Optimizing face validity, usability and quality of the web survey and TTO questions to ease understanding and increase engagement of the participants, using an iterative approach consisting of a qualitative study followed by the staged roll-out of the main survey

2) Developing an experimental design optimized for online data collection that is suited for a shorter survey length to maintain engagement of participants and to ensure cost-efficient sampling

3) Performing a sample size calculation to determine a sufficient sample size large enough to exclude poorer quality TTO data or to use only TTO data for generating the final value set

4) Investigating ways of assessing TTO data quality and then excluding TTO data with lower quality

Methods and survey development

Web survey development

We implemented the survey in HTML, CSS, JavaScript, PHP[11], twig [12], mariadb [13], and sqlite [14], featuring a mobile-first responsive design. This framework enabled us to customize the web survey to a larger degree than existing commercial solutions while adhering to Swedish data privacy and safety regulations. After the qualitative face-validity survey, we rolled-out the survey in three stages to iteratively analyse the collected data and in parallel improve the survey, until we deemed the data quality data to be satisfactory (Fig 1).

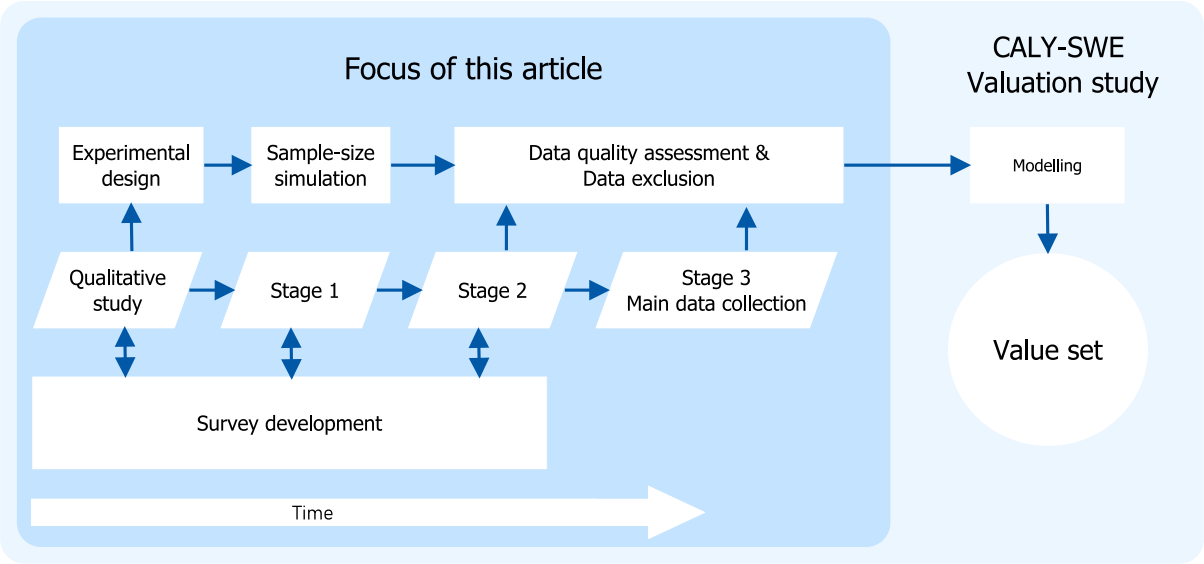

Fig 1. Conceptual flowchart of CALY-SWE valuation methods. ▱ symbolizes data collections and ▢ general processes and analysis steps. Arrows represent the workflow.

To improve usability, we focused on conveying information visually instead of relying on textual information. This included attribute levels of the states and the number of years for TTO questions. For both the DCE and TTO questions we used the same basic visual layout where attribute levels were visualized by adjacent vertical bars. The bars were filled according to the level of the 3 attribute levels, either empty, half, or full. Full CALY-SWE statement phrasing, and instructions were reachable over help buttons. For TTO questions we additionally displayed a horizontal bar that was filled according to the number of years in each state (See Supplementary, section on the web survey and Supplementary Fig S1 – S4).

Qualitative face validity study

To assess the face validity of the web survey, we conducted a qualitative study among 16 participants recruited from 2 local non-profit associations in Umeå, with varied age and gender. From September to December 2020, participants filled in the survey on a tablet at the premises of Umeå university and were afterwards interviewed by two

qualitative researchers (Jan Hjelte and Kerstin Edin) based on a semi-structured questionnaire.

Experimental design and sample size simulation

Our main consideration was to produce an experimental design with a relatively low number of questions per participant to keep the survey length short. To still collect enough data, we intended to compensate by increasing the number of participants. Because we aimed for a hybrid model, we also wanted to find the best balance between the number of TTO and DCE questions administered, and an adequate sample size for generating the value set with a TTO-only model.

To that end, we generated different experimental designs for different configurations of number of TTO and DCE questions. For each configuration, we chose the number of TTO and DCE question so that the predicted time for survey completion would be approximately 20 minutes, based on timing data from the qualitative study. Twenty minutes was the maximum we considered workable for cost-effectively sampling from the online panel, as costs per participant increase with increasing drop-out rates for longer surveys. The configurations consisted of 11 and 2, 8 and 3, 6 and 4, 4 and 5, and 3 and 6 DCE and TTO questions, respectively, with each 50, 100, 200, 300, 400, 500, 700, 1000, 1400, and 2000 participants.

To develop a D-optimal design for DCE, we used the orthogonal design approach presented in Street et al. [15]. We used this same design for all configurations and randomly picked the DCE questions for a block. For TTO, we generated the different designs specific to the number of TTO questions per configuration by using the *skpr* [16] package in R with the D-optimality criteria.

Thereafter, we ran simulations with different number of participants for configuration, using a hybrid and a TTO-data-only model, both with a varying intercept for the TTO part. We generated the data with the same varying intercept specification and parametrized it based on results from an earlier pilot study (unpublished). We evaluated the performance of the simulation in terms of the mean credible interval (CRI) width the mean absolute error of the weight posteriors compared to the generative model.

Comparison stage 1 and 2

Directly comparing data quality between stages was not possible due to low and differing sample sizes in each stage. We therefore applied a bootstrap approach to examine the effect of changes between the stages. We randomly sampled 500 DCE answers and 300 TTO answers 10000 times with repetition for each stage and performed a logit and an ordinary least squares linear regression for the bootstrapped DCE and TTO answers per stage with a main effects model. We then compared the distribution of the log likelihood, the mean standard error of the coefficients, and the number of inconsistent coefficients (level 2 greater than level 3) between stages.

TTO data quality and exclusion

The EQ-VT [9] focuses on interviewer training and poor quality data (defined by criteria such as if the interviewer explained the example long enough, or if the TTO answer for the pit state is the lowest TTO answer within a margin [17]) may be discarded. By design, self-administrated online surveys are unguided, making it impossible to rely on the interviewer performance for quality assessment.

To address this challenge, we explored alternative methods for assessing the quality of the TTO answers and for possibility excluding them. We developed a score based on inconsistent TTO answers that considers the severity of the inconsistencies, called the *combined inconsistency severity (CIS) score*.

To be logically consistent, participants' answers should value states with higher levels higher than those with lower levels for the same attribute. An *inconsistency* can occur for two TTO answers from a participant for the two states in a *dominated* choice. A dominated choice occurs if all attribute levels of one state are higher or equal while at least one level is higher than in the other state. If the TTO answer for the first state is lower or equal than for the second state, the two answers are not consistent with the dominated choice. Formally, for any two states S_j, S_k from a participant's TTO block with answers w_j ,w_k, and i indexing the attributes:

$$\forall i: S_{ji} \geq S_{ki} \wedge \exists i: S_{ji} > S_{ki} \wedge w_j \leq w_k, j \neq k$$

The *severity* of an inconsistency may be expressed as the absolute difference of the TTO answers (W) and the absolute difference in TTO level attributes for the two involved states (L), for participant p:

$$L_{pjk} = \sum_i \left| S_{ji} - S_{ki} \right|$$

$$W_{pjk} = \left| w_j - w_k \right|$$

For example, valuing 111111 (all six attributes on level 1, 'Do not agree') with 0.8 and 333333 (All six attributes on level 3, 'Agree completely') with 0.2 could be considered a quite severe inconsistency, with the level differences being 12 (L) and the weight difference (W) equalling 0.6. On the other hand, an inconsistency involving 232323 and 232333 with answers of 0.9 and 0.8 is less severe as there is only one level difference and a 0.1 weight difference. Such an inconsistency may occur due to the difficulty of the TTO question instead lack of engagement.

Across all values, we then normalized each score to [0,1] to make their scales comparable:

$$Lnorm_{pjk} = \frac{L_{pjk} - min(L_{pjk})}{max(L_{pjk}) - min(L_{pjk})}$$

$$Wnorm_{pjk} = \frac{W_{pjk} - min(W_{pjk})}{max(W_{pjk}) - min(W_{pjk})}$$

Afterwards, the score was summed for each participant p so that we could approximate data quality per participant and to not merely exclude single unfitting answers:

$$CISscore_p = \left(\sum_{jk} Lnorm_{pjk} + \sum_{jk} Wnorm_{pjk} \right)$$

As the TTO experimental design did not contain the same number of support points for possible inconsistencies for each participant (ranging from 5 to 7) and different blocks may have been more challenging to answer consistently, we normalized the scores per block. At last, we calculated the score percentile for each participant so that specific proportions of data can be excluded:

$$CISpercentilescore_p = \frac{rank_{block}(score_p)}{n_{block}}$$

We assessed the impact of including only participants with better quality TTO data by comparing the results of an ordinary linear regression for including 30% to 100% of the data according to the CIS score, in 1% increments. We specifically looked at the weights for 333333 and 111111 and their difference that defines the value set's range, the adjusted R^2, the number of inconsistent coefficients (where the coefficient for level 2 is larger than for level 3), the mean standard error and mean t-score of the coefficients (coefficients divided by the standard error) and the coefficient with their 95% confidence intervals.

We used R [18] for all data analyses with base R regression models except for the sample simulation with Bayesian hybrid models that were estimated with stan [19] and the cmdrstan R interface [20].

Results

Qualitative face validity study

Participants' statements revealed that the DCE and TTO questions were challenging but meaningful, partially engaging, and thought-provoking. Some stated that answering the DCE block before the TTO block helped them to increase their familiarity with the statement phrasing before tackling the more complex TTO questions. Most participants were able to independently finish the survey by just relying on the on-screen instructions.

During the qualitative interviews, we continuously developed the survey based on participant's feedback. For example, we added an explanation video for the TTO question, revised wording, and overhauled the DCE question layout to be in the same style as the TTO question layout. See Supplementary Table S3 for a detailed list of changes and differences between survey versions.

Experimental design sample size simulation

We based the DCE design on an orthogonal array with 6 columns and 45 rows, resulting in 43 pairwise comparisons after the removal of two dominated comparisons. We used random blocking where the DCE questions for each participant were randomly picked from the 43 comparisons.

For TTO, we deemed eight blocks with three states per block to be adequate, and we augmented each block with the pit state 111111 (all six attributes on level 1) so that 111111 would be evaluated by each participant, enabling to estimate 111111 with greater precision. The DCE and TTO designs are provided in Supplementary Tables S1 and S2.

The sample size simulation resulted in a decrease of the mean 95% CRI and mean absolute errors with increasing sample sizes. The hybrid model generally performed better than the TTO only model, especially for large proportions of TTO questions, but differences were small for both mean CRI width and mean absolute errors for configurations with at least four TTO questions. We decided for the configuration with four TTO questions and six DCE questions as we judged this configuration to offer a good balance of TTO and DCE questions. With this configuration, we found that at least 500 participants for using both TTO and DCE data and 1000 participants for only TTO data would be necessary. We decided to set the target sample size at 1500 participants, to have a safety margin and to leave the option to only use TTO data for generating the value set or to exclude poor quality data (Supplementary Fig S5).

Comparison stage 1 and 2

We collected data in 3 stages **(**Fig 1**)**: Stage 1 targeted 100 participants and took place from November 22 to December 2 2021. Based on preliminarily analysis after stage 1, we aimed to further improve the TTO question format and piloted the changes in Stage 2, targeting 200 participants (January 3 until January 12 2022). The main data collection in Stage 3 targeted 1500 participants (March 7 until April 18 2022).

In stage 2, we changed the iteration procedure so that participants choosing "equal" in the TTO question would not immediately proceed to the next question, but instead the interval of reachable values would shrink around the current bisection point, and renamed the option to "about equal". The goal was to reduce the incentive to finish the question faster by choosing equal. We further decreased the TTO time frame from 20 years to 10 years to reduce the expected number of needed iterations.

To facilitate understanding and increase engagement, we changed the graphical layout: Instead of selecting radio buttons, clicking on buttons would now directly submit the answer. We also introduced colour coding for the choices where the first and second choice were coloured in two distinct but neutral colours across DCE and TTO questions, randomized per participant (Supplementary Fig S2, S3). As a demonstrated example we also introduced a *learning* state as the first TTO state to showcase the trade-off mechanism (Details in Supplementary section on TTO learning state) and removed one DCE question per block to compensate the longer duration. Other changes included that participants could now navigate backwards. Supplementary Table S3 depicts a detailed list of changes.

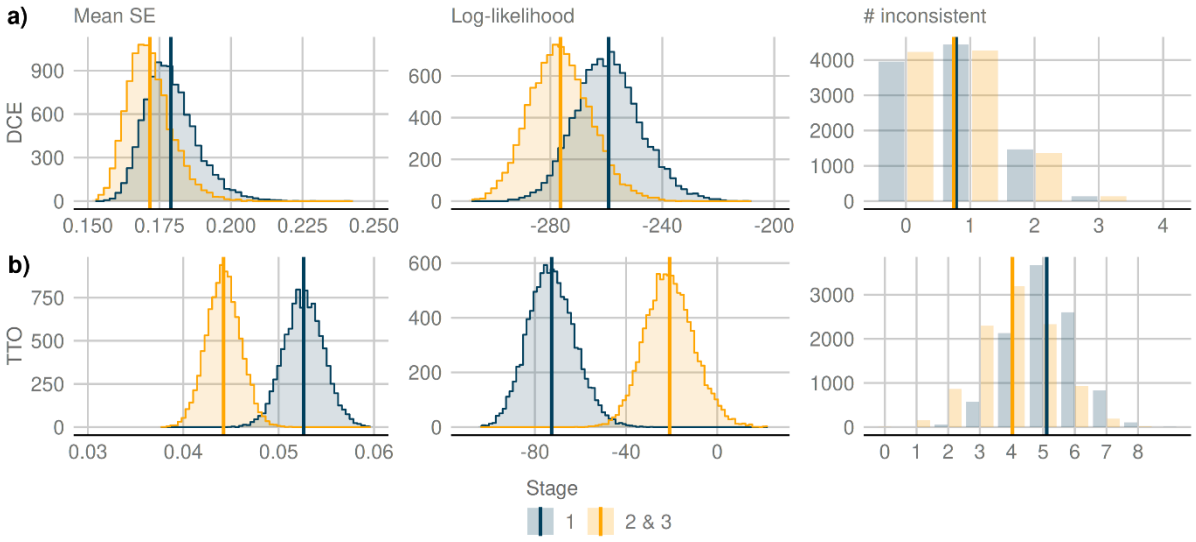

Fig 2. Histograms of results of bootstrapped comparison between stage 1 and 2.
a) Logistic regression of DCE data with 500 draws with repetition.
b) Linear regression of TTO data with 300 draws with repetition.
10000 bootstrap runs, main effects model. Discrete choice experiments (DCE). Time trade-off (TTO). Standard error (SE). Time trade-off (TTO).

The bootstrapped regressions from stage 2 compared to stage 1 indicated improvements in TTO data quality in terms of lower mean standard deviation and higher log likelihood with distinct distributions. The number of inconsistent coefficients also decreased on average but the distributions where overlapping. In comparison, the DCE distributions did not differ clearly between stages for all the indicators and indicated on average a decrease in data quality in stage 2 compared to stage 1. We deemed the data quality satisfactory and launched stage 3 without any further substantial changes. The final dataset included 199 stage 2 and 1498 stage 3 answers but no stage 1 data because of the survey differences.

TTO data quality and exclusion

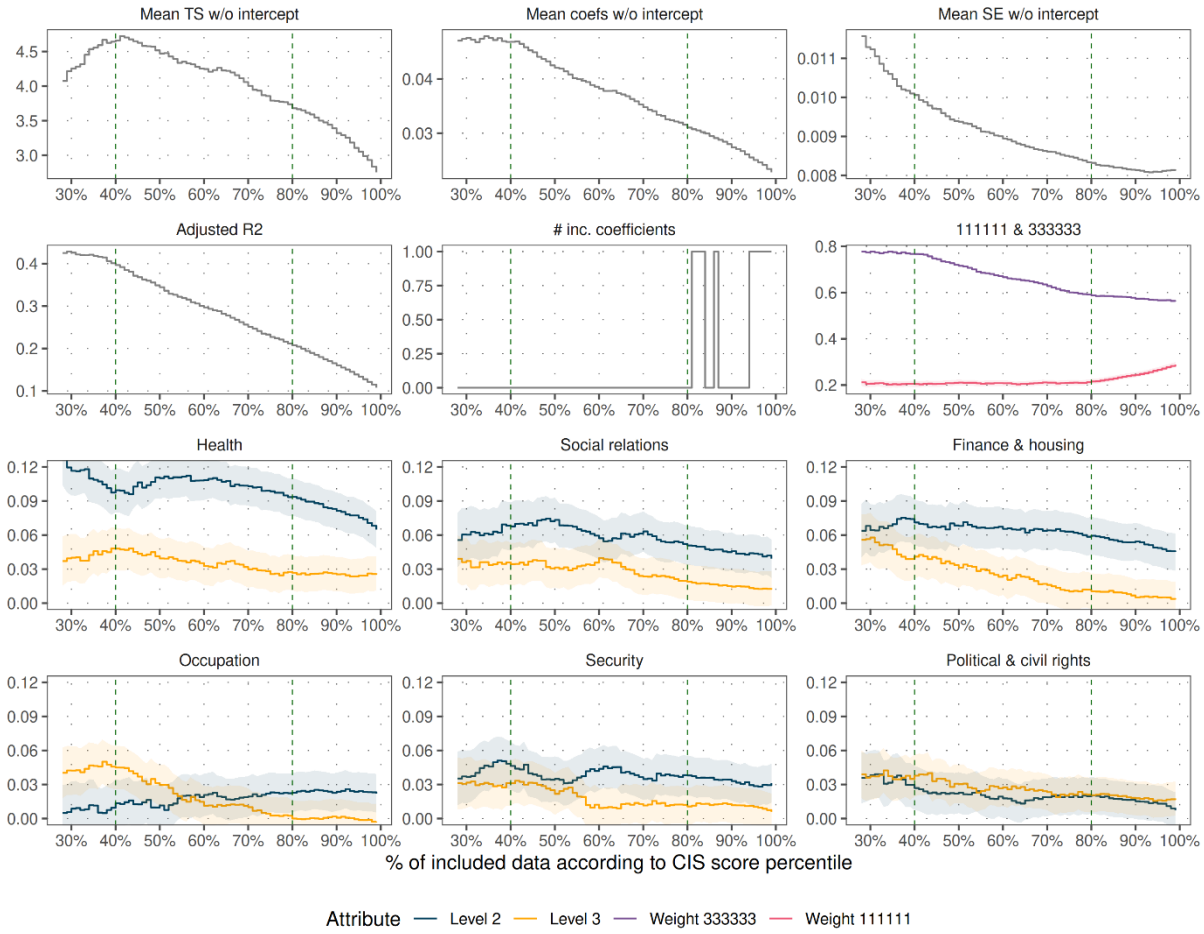

Fig 3. Line plots of effects of including participants according to the CIS score percentile, in 1% increments, in terms of the results of a linear regression main effects model. Shaded error bands represent 95% confidence intervals. T-score (TS). Standard error (SE). Combined inconsistency severity (CIS).

For higher percentages of included participants (100% corresponding to 1694) according to TTO CIS percentiles, the amount of variation explained by the model decreased. This was indicated by a decrease in the adjusted R^2, but also by a decrease of the mean t scores and by a general decrease of the coefficients and the range of

333333 to 111111. Put differently, the model was less able to systematically pick up inference patterns related to trade-offs between dimensions the more data was included. Instead, the fitted values converged towards the mean of the TTO answers: The intercept increased if more than 80% of data was included, together with a decrease of the level-attribute coefficients. Beyond including 80% the coefficient for occupation level 3 became increasingly inconsistent, likely related to the general decrease in coefficient magnitude. Interestingly, while in general the proportion of level 2 to level 3 coefficients remained constant, for occupation level 2 became larger than level 3 and for finance & housing the level 3 decreased compared to level 2 the more TTO data was included.

Therefore, the results suggest using between 40% to 80% of TTO data for generating the value set. Beyond 80%, the increase of the intercept would affect the value of 111111 and thus the lower anchoring of the value set on the [0,1] scale. In addition, the effect of the attribute level coefficients increasingly loses strength beyond 80%. Model performance of jointly estimated hybrid models with DCE data as well as representativity considerations may also play a role when deciding the exact amount of TTO data for generating the value set.

Discussion

We created an online-administered web survey for collecting TTO and DCE data for a CALY-SWE value set including an experimental design and we estimated the necessary sample size. We leveraged qualitative interviews and a two-stage survey roll-out to improve the web survey and developed a novel and sensitive way of excluding TTO data with poorer quality This work paved the way for eliciting a CALY-SWE value set which will be reported elsewhere.

Strengths include that we addressed the methodological challenges connected to online administration with an integrated approach that resulted in multiple benefits: 1) The qualitative face validity study allowed us to fine tune instructions and appearance of the web survey, and we were able to develop a web survey that focused on visually conveying the task instead of textual information or oral guidance by an interviewer. 2) Timing data from the qualitative study informed the sample size simulation which in turn facilitated an informed decision on the balance between the number of TTO and DCE questions and the sample size, enabling cost-efficient online sampling due to shorter survey length that is compensated by an increased sample size to collect enough data 4) The staged roll-out allowed to further improve the TTO data quality in stage 2 as indicated by the results of the bootstrapped regression analysis. 5) We developed a per-participant score that reflects the severity of TTO inconsistencies and enabled the exclusion of TTO data of an explicitly chosen proportion of participants to partially offset limited data quality connected to online self-administration. The CIS score also enable quality control for TTO data that is independent of the administration mode as it does not depend on interviewer performance.

Limitations include that we focused on a main effects only experimental design, thus not considering interactions. For this valuation study we focused on method development and producing a readily interpretable value set. For generating the DCE experimental designs we relied on orthogonal arrays as it was readily implementable without relying on additional software, but it may be less efficient compared to Bayesian experimental designs [21, 22].

Excluding TTO data based on inconsistencies is a contentious issue. In EQ-5D valuation studies sometimes evidently invalid answers such as always responding with the same value or the same pattern, or those selected in the feedback module, are excluded [23–26]. Excluding TTO data based on inconsistencies via the CIS score indeed constitutes a risk for data curation, were data is made to corresponds to prior norms because participants with larger uncertainties are discriminated against [27]. Similarly, concern over consistency of preferences over the range of excluded data, while generally stable in our study, still constitutes an important limitation as indicated by the changing order of coefficients for level 2 and 3 of occupation. On the other hand, under the assumption that a portion of the participants did not engage in the task sufficiently to correctly state their preferences, the score made it possible to transparently examine the impact of excluding different amounts of data, on a continuous range. Hence an informed, normative decision about how much data should be included could be made, and the data quality itself is defined in relation to the total sample. In contrast, excluding for example all participants that valued all states at 0.5 [9] is an absolute either-or decision without granular control over how much data is excluded.

Compromised TTO data quality resulting from online administration, compared to face-to-face data, or resulting from one interviewer per group compared one interviewer per person, has been previously quantified in the form of increased standard deviations [28], a lower number of trading iterations [10, 29], and a smaller range of the resulting value sets [10]. Those results align with our findings: Including poorer quality TTO data according to the CIS score resulted in a lower range and a larger standard error of the resulting weights. Similarly, changing the TTO iteration procedure in stage 2, so that choosing ‘equal’ still required additional iterations to arrive at an answer, improved the TTO data quality.

Other approaches to increase feasibility of valuation studies include videoconferencing which was found to be a viable alternative to face-to-face interviews for an Italian EQ-5D 5L valuation study [30]. Lipman similarly found that tele-TTO interviews were feasible [31]. Compared to online self-administration, this still requires interviewers and training to conduct the interview, imposing significant costs.

Further research into refinements of self-administered online valuation surveys may additionally increase their feasibility. This particularly includes instructions and visual design properties of TTO questions, and ways to assess data quality, but also to general knowledge on the feasibility of online administration in combination with TTO data quality assessment.

Conclusion

TTO and DCE data collected in self-administered online surveys may be used to elicit value sets if the web survey development, the experimental design, and sample size considerations are optimized and well-coordinated. The severity of inconsistencies can be used on a per-participant basis to identify and exclude poor quality TTO data that does not contribute to the modelling of preferences.

References

1. Månsdotter, A., Ekman, B., Feldman, I., Hagberg, L., Hurtig, A.-K., & Lindholm, L. (2017). We Propose a Novel Measure for Social Welfare and Public Health: Capability-Adjusted Life-Years, CALYs. *Applied Health Economics and Health Policy*, *15*(4), 437–440. https://doi.org/10.1007/s40258-017-0323-0
2. Meili, K. W., Månsdotter, A., Sundberg, L. R., Hjelte, J., & Lindholm, L. (2022). An initiative to develop capability-adjusted life years in Sweden (CALY-SWE): Selecting capabilities with a Delphi panel and developing the questionnaire. *PLOS ONE*, *17*(2), e0263231. https://doi.org/10.1371/journal.pone.0263231
3. Brazier, J., Ratcliffe, J., Salomon, J., & Tsuchiya, A. (2016). *Measuring and Valuing Health Benefits for Economic Evaluation* (Second Edition.). Oxford, New York: Oxford University Press.
4. VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proceedings of the National Academy of Sciences*, *114*(31), 8148–8156. https://doi.org/10.1073/pnas.1702996114
5. Lugnér, A. K., & Krabbe, P. F. M. (2020). An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. *Expert Review of Pharmacoeconomics & Outcomes Research*, *20*(4), 331–342. https://doi.org/10.1080/14737167.2020.1779062
6. Ternent, L., & Tsuchiya, A. (2013). A Note on the Expected Biases in Conventional Iterative Health State Valuation Protocols: *Medical Decision Making*. https://doi.org/10.1177/0272989X12475093

7. Rowen, D., Brazier, J., & Van Hout, B. (2014). A Comparison of Methods for Converting DCE Values onto the Full Health-Dead QALY Scale. *Medical Decision Making*, *35*(3), 328–340. https://doi.org/10.1177/0272989x14559542
8. Ramos Goñi, J. M., Craig, B. M., Oppe, M., & Van Hout, B. (2016). *Combining continuous and dichotomous responses in a hybrid model*. EuroQol Working Paper Series.
9. Stolk, E., Ludwig, K., Rand, K., van Hout, B., & Ramos-Goñi, J. M. (2019). Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health*, *22*(1), 23–30. https://doi.org/10.1016/j.jval.2018.05.010
10. Jiang, R., Shaw, J., Mühlbacher, A., Lee, T. A., Walton, S., Kohlmann, T., Norman, R., & Pickard, A. S. (2021). Comparison of online and face-to-face valuation of the EQ-5D-5L using composite time trade-off. *Quality of Life Research*, *30*(5), 1433–1444. https://doi.org/10.1007/s11136-020-02712-1
11. The PHP Group. (2022, September 11). PHP: Hypertext Preprocessor. Retrieved September 11, 2022, from http://php.net/
12. Symphony Project. (n.d.). Home - Twig - The flexible, fast, and secure PHP template engine. Retrieved September 21, 2022, from https://twig.symfony.com/
13. The Maria DB foundation. (2022, November 11). Maria DB. Retrieved November 16, 2022, from https://mariadb.org/about/
14. The SQlite project. (2022, November 11). SQLite. Retrieved November 16, 2022, from https://www.sqlite.org

15. Street, D. J., Burgess, L., & Louviere, J. J. (2005). Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, *22*(4), 459–470. https://doi.org/10.1016/j.ijresmar.2005.09.003
16. Morgan-Wall, T., & Khoury, G. (2021). Optimal Design Generation and Power Evaluation R: The skpr Package. *Journal of Statistical Software*, *99*(1). https://doi.org/10.18637/jss.v099.i01
17. Ramos-Goñi, J. M., Oppe, M., Slaap, B., Busschbach, J. J. V., & Stolk, E. (2017). Quality Control Process for EQ-5D-5L Valuation Studies. *Value in Health*, *20*(3), 466–473. https://doi.org/10.1016/j.jval.2016.10.012
18. R Core Team. (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
19. Stan Development Team. (2022). Stan. Retrieved from https://mc-stan.org
20. Gabry, J., Češnovar, R., Bales, B., Morris, M., Popov, M., Lawrence, M., Landau, W. M., & Socolar, J. (2022, August 11). cmdstanr. Retrieved from https://mc-stan.org/cmdstanr
21. Kessels, R., Jones, B., Goos, P., & Vandebroek, M. (2011). The usefulness of Bayesian optimal designs for discrete choice experiments. *Applied Stochastic Models in Business and Industry*, *27*(3), 173–188. https://doi.org/10.1002/asmb.906
22. Reed Johnson, F., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., Bresnahan, B. W., Kanninen, B., & Bridges, J. F. P. (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. https://doi.org/10.1016/j.jval.2012.08.2223

23. Devlin, N. J., Shah, K. K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics*, *27*(1), 7–22. https://doi.org/10.1002/hec.3564

24. Ludwig, K., Graf von der Schulenburg, J.-M., & Greiner, W. (2018). German Value Set for the EQ-5D-5L. *PharmacoEconomics*, *36*(6), 663–674. https://doi.org/10.1007/s40273-018-0615-8

25. Omelyanovskiy, V. (2021). Valuation of the EQ-5D-3L in Russia. *Quality of Life Research*, 11.

26. Welie, A. G., Gebretekle, G. B., Stolk, E., Mukuria, C., Krahn, M. D., Enquoselassie, F., & Fenta, T. G. (2020). Valuing Health State: An EQ-5D-5L Value Set for Ethiopians. *Value in Health Regional Issues*, *22*, 7–14. https://doi.org/10.1016/j.vhri.2019.08.475

27. Viney, R., Mulhern, B., Norman, R., Shah, K., & Devlin, N. (n.d.). Quality control vs. 'data curation': where should we draw the line in researcher judgements about.

28. Norman, R., King, M. T., Clarke, D., Viney, R., Cronin, P., & Street, D. (2010). Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. *Quality of Life Research*, *19*(4), 499–508. https://doi.org/10.1007/s11136-010-9609-5

29. Shah, K. K., Lloyd, A., Oppe, M., & Devlin, N. J. (2013). One-to-one versus group setting for conducting computer-assisted TTO studies: findings from pilot studies in England and the Netherlands. *The European Journal of Health Economics*, *14*(1), 65–73. https://doi.org/10.1007/s10198-013-0509-9

30. Finch, A. P., Meregaglia, M., Ciani, O., Roudijk, B., & Jommi, C. (2022). An EQ-5D-5L value set for Italy using videoconferencing interviews and feasibility of a new mode of administration. *Social Science & Medicine*, *292*, 114519. https://doi.org/10.1016/j.socscimed.2021.114519

31. Lipman, S. A. (2021). Time for Tele-TTO? Lessons Learned From Digital Interviewer-Assisted Time Trade-Off Data Collection. *The Patient - Patient-Centered Outcomes Research*, *14*(5), 459–469. https://doi.org/10.1007/s40271-020-00490-z

Declarations and statements

Author contributions (CRediT):

KWM: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, visualization, data curation, writing - original draft, writing – review and editing.

BM: Conceptualization, methodology, supervision, writing – review and editing

RS: Conceptualization, methodology, formal analysis, writing – review and editing

JH: Conceptualization, formal analysis, investigation, methodology, supervision, writing – review and editing

KE: Conceptualization, formal analysis, investigation, methodology, writing- review and editing

FN: Supervision, writing – review and editing

IF: Funding acquisition, Supervision, writing – review and editing

AM: Funding acquisition, Supervision, writing – review and editing

LL: Funding acquisition, Project administration, Supervision, writing – review and editing

Funding: Forte – Swedish Research Council for Health, Working Life, and Welfare, Grant No 2018-00143, Principal investigator Lars Lindholm

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Swedish Ethical Review Authority (Dnr 2021-04465).

Consent to participate: Informed consent was obtained from all individual participants included in the study, including the consent to publish results and anonymized data.