**Preference-Based Assessments**

**Modifying the Composite Time Trade-Off Method to Improve Its Discriminatory Power**

Michał Jakubczyk, PhD, Stefan A. Lipman, PhD, Bram Roudijk, PhD, Richard Norman, PhD, Eleanor Pullenayegum, PhD, Yaling Yang, PhD, Ning Yan Gu, PhD, Elly Stolk, PhD

**ABSTRACT**

**Objectives:** In cost-effectiveness analysis of health technologies, health state utilities are needed. They are often elicited with a composite time trade-off (cTTO) method, particularly for the widely used EQ-5D-5L. Unfortunately, cTTO discriminatory power is hindered by (1) respondents’ nontrading (NT) of time for quality, (2) censoring of utilities at $-1$, and (3) poor correlation of negative utilities with state severity. We investigated whether modifying cTTO can mitigate these effects.

**Methods:** We interviewed online 478 students (February to April, 2021) who each valued the same 10 EQ-5D-5L health states in 1 of 3 arms. Arm A used a standard cTTO, expanded with 2 questions to explore reasons for NT and censoring. Arms B and C used a time trade-off with modified alternatives offered to overcome loss aversion, to unify the tasks for positive and negative utilities, and to enable eliciting utilities $< -1$.

**Results:** In arms B and C, we observed less NT than in A (respectively, 4% and 4% vs 10%), more strictly negative utilities (38% and 40% vs 25%), and more utilities $\leq -1$ (18% and 30% vs 10%). The average utility of state 55555 dropped to $-2.15$ and $-2.52$ from $-0.53$; enabling finer trades in arm A reduced NT by 70%. Arm B and C yielded an intuitive association between negative utilities and state severity. These arms were considered more difficult and resulted in more inconsistencies.

**Conclusions:** The discriminatory power of cTTO can be improved, but it may require increasing the difficulty of the task. The standard cTTO may overestimate the utilities, especially of severe states.

**Keywords:** health state utility values, lexicographic preferences, loss aversion, time trade-off, worse than dead states.

**VALUE HEALTH.** 2022; 1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—1—
for examples). Therefore, the performance of cTTO is crucial for a reliable measurement of societal preferences for health. Unfortunately, cTTO still has shortcomings. In this article, we focus on its limited discriminatory power; that is, limited ability to capture the utility differences between health states, for very mild or severe states. For mild states, at the top of the utility scale, cTTO experiments occasionally end with nontrading (NT), that is, respondents not accepting any shortening of duration as equivalent to the worsening of QoL. In Golicki et al.6 NT occurred in > 6% EQ-5D-5L valuations, the proportion exceeded 30% for multiple mild health states, and > 50% respondents refused to trade for at least 1 state. NT results in equating the utility of Q and FH, which violates the logical ordering based on dominance (impaired state Q should have utility lower than $1 = u(FH)$).

Understanding the causes of NT, and perhaps reducing its frequency, is important for several reasons. If NT accurately reflects the preferences being based on lexicographic ordering (Qol only matters if longevity is equal), it implies that the standard QALY model does not hold (for more information on lexicographic frequency, is important for several reasons. If NT accurately re-

Methods

Possible Reasons for NT and AIT

We used 3 study arms—A, B, and C—designed following a set of hypotheses regarding possible reasons behind NT/AIT in the standard cTTO as presented below (other than simply reflecting the true preferences).

Regarding NT, it may trivially be caused by the limited granularity of responses in the standard cTTO, that is, minimum half-a-year steps. Alternatively, NT may stem from the following psychological phenomena. The cTTO starts with a comparison between living in FH for T years and living in Q for 10 years. Henceforth, for brevity, such health profiles are denoted by (FH, T) and (Q, 10), respectively. Initially, $T = 10$, and then $T$ is modified: if the former (latter) alternative is preferred, $T$ is reduced (increased), to worsen (improve) the first alternative. In this iterative process, the choices are framed as giving up time in Q to gain QoL. The confirmatory message in the EQ-VT protocol after reaching the indifference says: “[…] to avoid being in [Q] for 10 years you are willing to give up […] years out of 10. Is that correct?.” Such a framing presents the QoL change as a gain (“avoid being,” ie, having FH instead of Q) and the duration change as a loss (“willing to give up,” ie, having $T<10$ instead of 10). Accounting for loss aversion, a phenomenon well studied in economics17 and in health valuation more specifically,18 this framing magnifies the disutility of T reductions and reduces the propensity to trade (ie, give up time).

What adds to the loss aversion is the imprecision of preferences.19 The above-presented framing makes (Q, 10) the status quo, especially that it is a constant alternative in the BTD part. Foregoing status quo for (FH, T) requires the respondent to be strongly convinced. In consequence, the imprecise perception of the utility of (Q, 10) will lead to only accepting large T in (FH, T). Zhao and Kling20 used a similar reasoning to explain the willingness-to-pay/accept disparity.

Regarding AIT, if it results from true preferences, then assessing the actual utility values would be informative to correctly construct value sets. Alternatively, AIT may be enhanced by the discontinuity of preferences in the following sense. In LT-TTO, (FH, T) is compared with a mixed profile of (FH, 10) followed by (Q, 10), denoted henceforth with the “+” symbol as (FH, 10) + (Q, 10). The state Q only prevails in one of the alternatives. If a respondent derives disutility from the sole fact of having to experience T, this framing magnifies the disutility of T reductions and reduces the propensity to trade (ie, give up time).

Study Arms

This experiment used a design with 3 arms described below.

Arm A

The standard cTTO was used as implemented in the EQ-VT protocol predominantly used for the valuation of EQ-5D-5L (see Stolk et al.4 for a detailed description).

This arm started with a comparison (FH, 10) versus (Q, 10). Typically, the former was preferred. Then, a task separating BTD from WTD scenarios was used (to test whether $u(Q) \leq 0$:...
immediate death versus \((Q, 10)\). If \(Q\) was BTD and the latter was preferred, a single bisection step was used: \((FH, T)\) for \(T = 5\) versus \((Q, 10)\). Subsequently, the former was worsened if it was preferred (by reducing \(T\)) or improved otherwise. One-year steps were used, unless a change in direction was needed or near the corner comparisons \((T = 10\) or \(T = 0)\).

In the WTD case, 10 years were added to both lives to enable further trading. Hence, the following task was presented: \((FH, T)\) for \(T = 10\) versus \((FH, 10) + (Q, 10)\). This task effectively repeats the \(u(Q) \leq 0\) test. If the former was preferred, \(T\) was set to 5, and then 1-year steps were used (reduced to half a year in some cases, as for BTD). No further trading beyond \(T = 0\), that is, immediate death, versus \((FH, 10) + (Q, 10)\) was possible. In Figure 1, we present what the standard cTTO arm looked like.

After the completion of cTTO for all the states, additional questions were asked in arm A for states for which either NT or AIT occurred. In the former case, the respondents compared 2 hypothetical worsenings (to make them both losses): of QoL from FH to Q or of the duration from 10 years (in months, weeks, or days), to indicate when they would be indifferent.

For states ending in AIT, the respondents compared \((FH, 5) + (Q, 5)\) with \((FH, 10) + (Q, 10)\), to check whether \(u(Q)\) is \(< -1\), \(-1\), or \(> -1\), while including \(Q\) in both alternatives in case of discontinuous preferences. The exact wording of added questions is presented in the Appendix A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.08.011.

**Arm B**

Arm B was designed to avoid loss aversion and include \(Q\) in both alternatives in the WTD scenario (to avoid the impact of preference discontinuity). The first choice task was as in arm A. However, if FH was preferred, \((Q, 10)\) was improved by adding \((FH, 10)\) at the beginning. Hence, in the BTD scenario, \((FH, 10)\) was compared with \((FH, T) + (Q, 10), 0 \leq T \leq 10,\) and \(T\) was increased (decreased) if the first (second) alternative was preferred.

---

**Figure 1.** Arm A, examples of choice tasks for better than dead (upper) and worse than dead (lower) scenario. The choice task separating the 2 scenarios in the middle (testing \(u(Q) \leq 0\)). The indifference would be interpreted as \(u(Q) = 0.5, 0,\) and \(-0.5,\) from top to bottom.
resemble arm A, the first 3 choice tasks corresponded to \( u(Q) = 1, 0, \) and 0.5. Then, 1-year steps were used (or half a year, as in arm A).

In the WTD part, to include Q in both lives, \( (FH, T) + (Q, 10 - T) \) was compared with a constant alternative \( (FH, 10) + (Q, 10) \). Initially, \( T = 10 \) in the BTD-WTD separating task, and subsequently \( 0.5 \leq T \leq 9.5 \). Because the minimal elicitable utility value amounted to as little as \( u(Q) = -19 \) (for \( T = 0.5 \)), no bisection was used, but instead \( T \) was changed in 1-year steps from \( T = 10 \) (reduced to half a year, as in arm A). The same number of unique \( T \) values is available in arms B and A for BTD part, and an almost identical number for WTD part (nevertheless, a wider range of negative utilities reduces the granularity). Arm B is illustrated in Figure 2.

Arm C was designed to avoid a change of framing between the BTD and WTD scenarios. In this unified framework, the respondent compared \( (FH, 10 - 2T) + (Q, T) \) with \( (FH, 2T) + (Q, 10 - T) \), starting with \( T = 0 \), that is, \( (FH, 10) \) versus \( (Q, 10) \) just like in arms A and B. The BTD-WTD separating choice task is obtained for \( T = 2.5 \): \( (FH, 5) + (Q, 2.5) \) versus \( (FH, 5) + (Q, 7.5) \), in which the alternatives differ only in time spent in Q. Increasing \( T \) makes the first alternative less attractive and the second alternative more attractive. Half-a-year changes in \( T \) were used, \( 0 \leq T \leq 4.5 \). Because only 10 unique values of \( T \) are available, no bisection was used in either BTD or WTD scenario. The minimal elicitable utility amounts to \(-8\). The choice tasks are illustrated in Figure 3.
Interviewing

A convenience sample of students were randomized between the arms and interviewed by 5 interviewers online with screen sharing and a web application. More details can be found in the Appendix B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.08.011.

Interviews started with a welcome and some general information. Then, the respondents were asked about demographics, own health, and experience with health problems. The main part of the interview consisted of 3 warm-up tasks (living in a wheelchair and mild, severe, and then random order): (mild) 11121, 11211, 12111, and 21111; (moderate) 22222 and 33333; and (severe) 43555, 44444, 55424, and 55555, based on the data on the states with most NT and AIT in the Dutch valuation.
Table 1. Sample structure: number of interviews included, sex, age (mean and SD), own health as per VAS and EQ-5D-5L, and experience with various health problems.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Arm A</th>
<th>Arm B</th>
<th>Arm C</th>
<th>Arm D</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>159</td>
<td>157</td>
<td>156</td>
<td>472</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>92 (58)</td>
<td>84 (54)</td>
<td>100 (64)</td>
<td>276 (59)</td>
</tr>
<tr>
<td>Mean age (SD), years</td>
<td>22.7 (3.2)</td>
<td>22.7 (3.2)</td>
<td>22.9 (3.1)</td>
<td>22.7 (3.2)</td>
</tr>
<tr>
<td>Mean VAS (SD)</td>
<td>81.9 (10.6)</td>
<td>81.7 (11.4)</td>
<td>82.5 (9.7)</td>
<td>82.0 (10.6)</td>
</tr>
<tr>
<td>11111, n (%)</td>
<td>54 (34)</td>
<td>52 (33)</td>
<td>62 (40)</td>
<td>168 (36)</td>
</tr>
<tr>
<td>11112, n (%)</td>
<td>22 (14)</td>
<td>28 (18)</td>
<td>32 (21)</td>
<td>82 (17)</td>
</tr>
<tr>
<td>11121, n (%)</td>
<td>22 (14)</td>
<td>15 (10)</td>
<td>14 (9)</td>
<td>51 (11)</td>
</tr>
<tr>
<td>11122, n (%)</td>
<td>16 (10)</td>
<td>14 (9)</td>
<td>9 (6)</td>
<td>39 (8)</td>
</tr>
<tr>
<td>11222, n (%)</td>
<td>3 (2)</td>
<td>8 (5)</td>
<td>7 (5)</td>
<td>18 (4)</td>
</tr>
<tr>
<td>11113, n (%)</td>
<td>8 (5)</td>
<td>4 (3)</td>
<td>4 (3)</td>
<td>16 (3)</td>
</tr>
<tr>
<td>11212, n (%)</td>
<td>6 (4)</td>
<td>5 (3)</td>
<td>2 (1)</td>
<td>13 (3)</td>
</tr>
<tr>
<td>Own health problems, n (%)</td>
<td>35 (22)</td>
<td>21 (13)</td>
<td>27 (17)</td>
<td>83 (18)</td>
</tr>
<tr>
<td>Family health problems, n (%)</td>
<td>126 (79)</td>
<td>125 (80)</td>
<td>110 (71)</td>
<td>361 (77)</td>
</tr>
<tr>
<td>Family premature death, n (%)</td>
<td>76 (48)</td>
<td>71 (45)</td>
<td>67 (43)</td>
<td>214 (45)</td>
</tr>
<tr>
<td>Religion, n (%)</td>
<td>31 (20)</td>
<td>31 (20)</td>
<td>31 (20)</td>
<td>93 (20)</td>
</tr>
</tbody>
</table>

Note. One person in arm C did not identify with any sex. Regarding own health, only health states prevailing in at least 10 respondents in the whole sample are reported. SD indicates standard deviation; VAS, visual analog scale.

Finally, debriefing questions focused on difficulty of the task were asked.

Analysis

We analyzed the elicited utilities per state and arm with cumulative distribution functions (cdf), means and standard deviations, and the proportion of NT, negative values, and ≤−1 values (comparisons used Cochran-Mantel-Haenszel test stratified by state). In arm A, we also studied the utilities elicited with additional questions. Because in arms B and C much lower negative utilities can be elicited, whether all years were traded in the WTD part of the arms is not comparable. Hence, no counterpart of AIT was studied in arms B and C.

We studied the correlation between negative utility and state severity. Admittedly, the slope in such a regression is biased toward 0 compared with the regression for all (not only negative) utility values.10 Studying the relationship is still informative to gauge the effect on the association between state severity and negative utilities of the change in the BTD-WTD separating task or the alternatives offered when eliciting negative utilities. To expand the existing approach of Gandhi et al,15 we used as a measure of severity (beside LSS) also the proportion of WTD and the mean positive utility Appendix E https://doi.org/10.1016/j.jval.2022.08.011. Given that utilities in arm A are censored at −1, we introduced the same censoring for arms B and C in this analysis for comparability. To inspect data quality, we checked for interviewer effects Appendix E https://doi.org/10.1016/j.jval.2022.08.011 and compared the arms in terms of inconsistencies (assigning greater utility to an objectively worse state). In each interview, there were 35 unique pairs of different, Pareto-ranked states. We calculated the proportion of pairs with an illogical valuation ordering.

We also regressed observed disutilities on LSS for each respondent separately, to detect problems with understanding. To compare the respondents’ behavior, we analyzed the time spent on the task, the number of steps required to reach a final value, and the subjective difficulty Appendix E https://doi.org/10.1016/j.jval.2022.08.011.

Results

Sample Characteristics

We collected 478 interviews in February to April 2021. Five interviews were removed because of internet problems and data loss, and 1 interview because of evident respondent inattentiveness (as judged by an interviewer). The demographic structure of the sample is presented in Table 1.

Utilities Per State and Arm

The main results are presented in Table 2 (means, standard deviations, and proportion of NT and u = −1) and illustrated in Appendix C in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.08.011. For all the states, the difference in means between arm A and each of the other arms was highly significant (P < .001, t test).

The experimental arms reduced the proportion of NT, with odds ratio (OR) amounting to 0.36 (95% confidence interval [CI] 0.26-0.48) for B versus A and 0.374 (95% CI 0.28-0.51) for C versus A. Conversely, the experimental arms increased the proportion of u = −1: OR = 2.53 (95% CI 1.99-3.21) for B versus A and OR = 8.04 (95% CI 6.27-10.33) for C versus A (all P < .001).

The experimental arms increased the proportion of states being WTD, u = 0 (or strictly WTD, u<0). In arms A to C, respectively, it amounted to 28%, 40%, and 44% (or 25%, 38%, and 40%), with the differences between A and the remaining arms being significant (P < .001).

NT and AIT in Additional Questions in Arm A

With finer trades allowed, the respondents traded in 70% of NT cases (see Table 3). For 22222, some respondents traded more than what is possible in cTTO. We attribute it to the erratic
behavior of some respondents combined with a very small sample for this state. Among all AIT cases in the standard cTTO part, 92% ended in the respondent strictly preferring immediate death, that is, demonstrating \( u(Q) < -1 \). This high proportion was confirmed in the additional question: the respondents mostly (77%) preferred \((FH, 5) + (Q, 5)\) over \((FH, 10) + (Q, 10)\), which also corresponds to \( u(Q) < -1 \), under the QALY model.

**The Overall Distribution Properties**

In Figure 4, we present the cdf per state and arm. In A, for severe states, there are many close-to-0 positive utility values and few close-to-0 negative values. As a result, the cdf have a kink around 0: they are flat in the range \((-0.5, -0.05)\), rise steeply in \((0, 0.3)\), and flatten again in \((0.3, 1)\). Such a pattern is not discernible for arms B and C: the cdf have a convex-concave shape (when smoothed and treated as continuous), as for typical probability distributions (eg, normal or logistic).

In Figure 5, we illustrate the association of negative utilities with 3 measures of state severity: LSS, proportion of respondents for whom the state was WTD, and the mean positive utility. For arm A, the findings of Gandhi et al\(^{15}\) are replicated (ie, no correlation) and extended to other measures of severity, whereas the correlation is present for arms B and C. The correlation is also present when only looking at the states that were considered WTD in arm A (so in this sense, it does not stem from the overall utility being decreased in arms B and C).

**Respondents’ Consistency**

For each respondent, the association of LSS with disutility was positive, as expected. The mean proportion of state pairs violating logical order amounted to 1.56%, 4.79%, and 2.69%, in respective arms (significantly lower in arm A: \( P < .001 \) vs B and \( P = .002 \) vs C). To account for the size of inaccuracy, for each interview we calculated the minimal utility shift required (ie, the aggregate absolute change in utility) that would allow to remove all the inconsistencies. The average distortion needed amounted to 0.09, 0.52, and 0.94 in respective arms \((P = .001 \text{ vs } B \text{ and } P < .001 \text{ vs } C)\). The increase in the proportion and size of inconsistencies in experimental arms agrees with perceived greater difficulty of the task (see Appendix E in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.08.011).

### Discussion

**The Ends of TTO**

Regarding the upper end of the utility scale, the cTTO modifications decreased the proportion of NT. This was true both of additional questions in arm A (offering finer trades) and the changed structure of choices in arms B and C. Therefore, the lexicographic preferences seem not to be as prevalent as the standard cTTO could suggest. As such preferences violate the QALY model, this is a convenient finding from the theoretical perspective. The 2 modifications of cTTO could be combined, to further reduce NT.

Arms B and C were designed not to invoke loss aversion. We observed more trading in these arms, which suggests that loss aversion is indeed present in standard cTTO. This finding complements earlier research by Lipman et al\(^{18}\) who showed that correcting for loss aversion (in an approach different to ours) also decreases the utility values. We see no other mechanisms that can explain our findings. For instance, discounting would not serve this purpose given that for small trades the choices in arms A and B are similar: for example, instead of comparing \((FH, 9.5)\) with \((Q, 10)\) in arm A, we have \((FH, 10)\) versus \((FH, 0.5) + (Q, 10)\) in arm B. We also do not think our findings can be explained by the ordering effect (ie, preference for the QoL improving rather than worsening over time). In arm C and in the WTD part of arm B, the ordering effect would cancel out between the 2 alternatives. In the BTD part of arm B, the aversion to QoL worsening over time would increase the elicited utility value.

### Table 2. Utility values per state and arm: mean, SD, % of NT and \( u \leq -1 \).

<table>
<thead>
<tr>
<th>State</th>
<th>Arm A mean (SD)</th>
<th>% NT</th>
<th>( u \leq -1 )</th>
<th>Arm B mean (SD)</th>
<th>% NT</th>
<th>( u \leq -1 )</th>
<th>Arm C mean (SD)</th>
<th>% NT</th>
<th>( u \leq -1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>11121</td>
<td>0.91 (0.11)</td>
<td>21</td>
<td>0</td>
<td>0.78 (0.26)</td>
<td>10</td>
<td>0</td>
<td>0.69 (0.24)</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>11211</td>
<td>0.91 (0.09)</td>
<td>23</td>
<td>0</td>
<td>0.77 (0.29)</td>
<td>10</td>
<td>1</td>
<td>0.70 (0.25)</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>12111</td>
<td>0.92 (0.11)</td>
<td>28</td>
<td>0</td>
<td>0.78 (0.23)</td>
<td>10</td>
<td>0</td>
<td>0.69 (0.29)</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>21111</td>
<td>0.91 (0.11)</td>
<td>26</td>
<td>0</td>
<td>0.76 (0.23)</td>
<td>9</td>
<td>0</td>
<td>0.69 (0.26)</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>22222</td>
<td>0.74 (0.25)</td>
<td>3</td>
<td>0</td>
<td>0.53 (0.43)</td>
<td>3</td>
<td>1</td>
<td>0.43 (0.78)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>33333</td>
<td>0.49 (0.36)</td>
<td>0</td>
<td>2</td>
<td>0.17 (0.59)</td>
<td>1</td>
<td>8</td>
<td>-0.26 (1.35)</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>43555</td>
<td>-0.40 (0.52)</td>
<td>0</td>
<td>23</td>
<td>-1.36 (3.14)</td>
<td>0</td>
<td>41</td>
<td>-2.14 (2.40)</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>44444</td>
<td>-0.29 (0.54)</td>
<td>0</td>
<td>20</td>
<td>-1.48 (3.72)</td>
<td>0</td>
<td>38</td>
<td>-1.81 (2.20)</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>55424</td>
<td>-0.21 (0.56)</td>
<td>0</td>
<td>20</td>
<td>-1.20 (2.94)</td>
<td>0</td>
<td>37</td>
<td>-1.55 (2.20)</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>55555</td>
<td>-0.53 (0.46)</td>
<td>0</td>
<td>30</td>
<td>-2.15 (4.27)</td>
<td>0</td>
<td>52</td>
<td>-2.52 (2.63)</td>
<td>0</td>
<td>79</td>
</tr>
</tbody>
</table>

NT indicates nontrading; SD standard deviation.

### Table 3. Arm A, trading when offered shorter intervals in additional question (time in days).

<table>
<thead>
<tr>
<th>State</th>
<th>n, % trading</th>
<th>Time traded: mean (SD), median</th>
<th>Time (nonzero): mean, median</th>
</tr>
</thead>
<tbody>
<tr>
<td>11121</td>
<td>33, 64%</td>
<td>50 (126), 7</td>
<td>78, 60</td>
</tr>
<tr>
<td>11211</td>
<td>37, 68%</td>
<td>52 (132), 7</td>
<td>77, 30</td>
</tr>
<tr>
<td>12111</td>
<td>44, 68%</td>
<td>42 (70), 8</td>
<td>61, 30</td>
</tr>
<tr>
<td>21111</td>
<td>42, 67%</td>
<td>49 (84), 12</td>
<td>73, 30</td>
</tr>
<tr>
<td>22222</td>
<td>5, 80%</td>
<td>393 (483), 150</td>
<td>492, 435</td>
</tr>
</tbody>
</table>
Figure 4. Cumulative distribution functions for elicited utilities (graph censored at $-3$ for clarity).
Figure 5. The association between the negative utility ($u^-$, censored at $-1$) and 3 measures of state severity: sum of dimension levels (LSS), proportion of valuations WTD, and mean positive utility ($u^+$). Results presented by arm (A-C, consecutive rows). Notice the difference in vertical scale (smaller for arm A); optically same slopes correspond to smaller actual values. As confirmed by the simulation results, negative associations are expected in the left 2 columns and a positive one in the right column (see Appendix D in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.08.011).

LSS indicates level sum score; WTD, worse than dead.
Regarding the lower end of the scale, had AIT in cTTO been caused by discontinuity of preferences, the utilities elicited with B and C should increase. Nevertheless, experimental arms increased the frequency of $u \leq -1$, providing no evidence to support our hypothesis on discontinuity.

An important conclusion from our findings is that the large number of AIT cases in the standard cTTO is not an artifact (ie, it is not caused by the design of cTTO) but it represents true preferences. This is confirmed by the answers of respondents to the additional question in arm A (see NT and AIT in Additional Questions in Arm A) and the elicited utilities in arms B and C being often $<-1$. In consequence, the standard cTTO is subject to a serious censoring problem, and the credibility with which the range of negative utilities is elicited is limited.

Perhaps some part of NT and AIT observed in valuation studies results from carelessness of some respondents. Then, it may not be removed by modifications of the protocol. Nevertheless, strict quality control used in the valuation studies helps to spot and reduce the impact of such problematic respondents.21

**About Dead**

We also observed important effects in the middle of the QALY utility scale. We found many more WTD states in arms B and C than in A. This is particularly striking given that arm C offered much coarser granularity of negative values, so it could be expected that fewer respondents would assign a strictly negative value. This finding is complemented by the comparison of the cdf between arms (see The Overall Distribution Properties).

We attribute these findings to a more seamless transition from the BTD to WTD part in the experimental arms, where the structure of choices changes little (especially in arm C, see Fig. 3). In arm A, to enter the WTD part, the respondent must choose immediate death over $Q$—the only comparison in cTTO (except for AIT case) that involves a 0-duration alternative. Even if the comparison versus immediate death may be theoretically appealing, given that it removes some confounding factors (for instance, discounting or QoL changing over time), it may be appalling to some respondents, because imagining immediate death may trigger strong emotions. Such a choice is qualitatively different from other cTTO choices.22,23

In consequence, the comparison versus immediate death in arm A works as a strong gatekeeper preventing respondents from declaring a state Q as WTD. In consequence, respondents who would attribute slightly negative $u(Q)$ under other elicitation methods may not pass the gatekeeper, causing clustering for slight positive utilities. Nevertheless, once $u(Q)$ is sufficiently negative to pass the gate, many years are traded off in the LT-TTO, resulting in the kink of cdf and distorting the correlation between the severity and negative utilities (especially in view of the floor effect in arm A, ie, censoring the utilities in $-1$). The abovementioned problems are apparently avoided in arms B and C. The evidence for respondents avoiding reporting $u(Q) \leq 0$ in the standard cTTO was also presented by Lipman et al.25

**Implications for Value Sets**

Our findings suggest that the standard cTTO may overestimate the utility over the whole scale and in particular for more severe states. This is not to say that arms B and C provide correct estimates, especially that these 2 arms differ between each other. Because this is definitely beyond the scope of this article, we leave aside the discussion about the meaning of “true value” of utility, given that our article clearly shows that the elicited values strongly depend on how TTO is operationalized. In addition, because arms B and C posed more difficulties to the respondents, we do not conclude they should replace the standard cTTO. Instead, the experimental arms highlight where the standard cTTO could fail.

This likely overestimation in arm A may explain some earlier findings. Wu et al26 conducted a large study in which 3320 respondents completed TTO and DCE with duration (DCEd) for health states defined using the SF-6Dv2 descriptive system. Health state utilities as elicited with TTO varied between $-0.28$ and 1 with 5% states valued as WTD. The utilities obtained by DCEd were lower, between $-0.54$ and 1, with 8.5% states WTD. The values were highly correlated between the approaches (intraclass correlation coefficient equal to 0.98), which suggests that the difference lies mostly in the scale assigned and not a change in preferences. An earlier pilot study by Xie et al27 anchors the latent utilities on the QALY scale.

We acknowledge that an opposite difference was observed when DCEd and TTO data were estimated in independent samples from the same general population. For instance, Craig and Rand28 report a US value set based on DCEd with $u(55555) = -0.29$ and Pickard et al4 a TTO-based value set with $u(55555) = -0.57$. Nevertheless, Craig and Rand28 accounted for discounting that makes the results less comparable, given that such correction has been shown to increase the estimated utilities.

Because of the overestimation of utilities, the health gains resulting from QoL improvements are underestimated and those resulting from life prolongation are overestimated. The low discriminatory power at the top of the utility scale may discriminate against public health interventions offering small QoL gains for small incremental costs.

**Limitations**

The very low negative utilities and averages for severe states are surprising and lack face validity (at least to those who are used to existing value sets obtained with existing methods). Using elicitation methods with no censoring creates the risk that some respondents will assign very low, negative utilities, even if due to misunderstanding of the task. Then, some careful procedures of removing extreme values should be implemented to avoid distortion in value set. The possibility of eliciting very negative values was considered a problem with previous approach to TTO, as used in the Measurement and Value of Health study by Dolan,30 which allowed eliciting utilities as low as $-39$. Because such very low utilities were observed, first a theoretically unfounded algebraic correction was introduced, to be subsequently replaced by cTTO. Nevertheless, our results show that limiting interest narrowly to the $[-1, 1]$ range is probably unwarranted.

In addition, the low values obtained for mild and moderate states in arms B and C of our study lack face validity. Some of these low values may stem from the erratic behavior of some respondents also resulting in increased inconsistency rate observed in these arms (the proportion of inconsistency is still not worrying compared with other studies25). That the experimental arms pose more difficulty to the respondent is also confirmed directly by the respondents in debriefing questions (see Appendix E in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.08.011). Think-aloud studies, assuming they provoke more deliberate choices, could provide additional insight into what extent such low utilities are credible and what thought processes can lead to them.

Another limitation is a very coarse granularity of possible answers in arm C. This was the result of trying simultaneously to (1) include both Q and FH in the alternatives, (2) avoid time steps $< 6$ months, and (3) not expand the total time horizon. Nevertheless, arm C gave similar results to B in many aspects (eg, the proportion of NT or how it restored the association between negative utility
and state severity). In this sense, arm C provided confirmatory value to the whole study. The very negative values in C may result from the small number of steps needed to reach these very negative values: although in cTTO after 3 steps we may reach the utility of −0.5, in arm C it can be \( u = -8 \).

We admit that considering nonconstant alternatives (i.e., ones in which QoL changes over time) is more cognitively demanding for the respondents. In cTTO no such alternatives are used in BTD, and in WTD only one of the alternatives is nonconstant. Nevertheless, in the experimental arms, often 2 alternatives were nonconstant. This fact might be the main reason for the increased difficulty and inconsistencies. Hence, in future research, attempts should be made to refine the standard cTTO but still refrain from nonconstant profiles as much as possible.

In the context of task difficulty, we used a sample that was highly nonrepresentative for the general population: only young and educated individuals. On the one hand, arms B and C might yield even more inconsistencies in less educated samples. On the other hand, perhaps some part of the difficulty resulted from the use of an online format. Because we did not aim to produce a value set, the lack of representativeness is not a limitation per se. In the convenience sample we used, we expect the same direction of effects of TTO modification, differing only in size.

Regarding this size, we largely surveyed the Dutch (even if a substantial fraction of the students from whom we recruited was international; alas, no data on nationality were collected), and more trading can typically be expected in Dutch samples.\(^{31}\)

We believe also that young people, typically not accustomed to health problems, may be more willing to trade to avoid them. Therefore, it would be interesting to see whether similar effects are observed in other samples.

Finally, our study focused on mild and severe states. Hence, the performance of experimental arms in the middle of the scale is more difficult to assess. Further studies could attempt at gathering more evidence.

Conclusions

The experimental TTO arms alleviate some issues with cTTO at the upper end of the utility scale (NT), middle of the scale (crowding at or above 0), and negative utilities (removing the censoring at −1 and counterintuitive patterns when related to state severity). This suggests that standard cTTO may overestimate the utilities. Nevertheless, the experimental arms increased inconsistencies and yielded utility values with problematic face validity. This suggests that further research is needed and it should focus on reducing the effect of loss aversion and determining whether the state is better or WTD in a way more compatible with other tasks.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.08.011.

Article and Author Information

Accepted for Publication: August 16, 2022

Published Online: xxxx
doi: https://doi.org/10.1016/j.jval.2022.08.011

Author Affiliations: Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland (Jakubczyk); Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands (Lipman); EuroQol Research Foundation, Rotterdam, The Netherlands (Roudijk, Stolk); School of Population Health, Curtin University, Perth, WA, Australia (Norman); Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada (Pullenayegum); Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, ON, Canada (Pullenayegum); Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, England, UK (Yang); School of Nursing and Health Professions, University of San Francisco, San Francisco, CA, USA (Gu).

Correspondence: Michal Jakubczyk, PhD, SGH Warsaw School of Economics, Al. Niepodleglosci 162, 02-554 Warsaw, Poland. Email: michal.jakubczyk@sgh.waw.pl

Author Contributions: Concept and design: Jakubczyk, Lipman, Roudijk, Norman, Pullenayegum, Yang, Gu, Stolk. Acquisition of data: Jakubczyk, Lipman, Roudijk. Analysis and interpretation of data: Jakubczyk, Lipman, Roudijk, Norman, Pullenayegum, Yang, Gu, Stolk. Drafting of the manuscript: Jakubczyk, Norman, Stolk. Critical revision of paper for important intellectual content: Jakubczyk, Lipman, Roudijk, Norman, Pullenayegum, Yang, Gu, Stolk. Statistical analysis: Jakubczyk. Obtaining funding: Jakubczyk, Lipman, Roudijk, Norman, Yang, Gu, Stolk. Administrative, technical, or logistic support: jakubczyk, Lipman, Roudijk. Supervision: Jakubczyk.

Conflict of Interest Disclosures: Dr Jakubczyk reported receiving grants from the EuroQol Research Foundation, during the conduct of the study, and grants and personal fees from the EuroQol Research Foundation, outside the submitted work. Drs Lipman, Pullenayegum, and Yang reported receiving grants from the EuroQol Research Foundation, during the conduct of the study, and grants from the EuroQol Research Foundation, outside the submitted work. Dr Roudijk reported receiving grants from and other financial relationships with the EuroQol Research Foundation, outside the submitted work. Dr Norman reported receiving grants from the EuroQol Research Foundation, during the conduct of the study, and grants and personal fees from the EuroQol Research Foundation, outside the submitted work. Dr Ning Yan Gu is a member of EuroQol Research Foundation and reported received grants from the EuroQol Research Foundation for this study. Dr Stolk reports employment at and grants from the EuroQol Research Foundation, outside the submitted work. The EuroQol Research Foundation is the sole owner of the EuroQol Valuation Technology, which is a software tool for valuing health. EuroQol also has the intellectual property rights on the modified versions of EuroQol Valuation Technology tested in this article. The views presented in this article may not be shared by the EuroQol Research Foundation or the EuroQol Group. Drs Norman and Stolk are editors for Value in Health and had no role in the peer-review process of this article.

Funding/Support: This work was supported by grant number EQ Project 20190070 from the EuroQol Research Foundation.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgment: The authors appreciate the funding of this study by the EuroQol Research Foundation. The views expressed by the authors in the paper do not necessarily reflect the views of the EuroQol Foundation. The authors thank Meike van der Linden, Ilse Littooy, and Angeliki Kaproulia for their help in data collection. They also appreciate the cooperation from HostLab when developing and using the software for interviews.

REFERENCES

14. Jakubczyk M. What if 0 is not equal to 0? Inter-personal health utilities anchoring using the largest health gains. Accepted for publication in *Eur J Health Econ*. 2022.