Received: 20 September 2019 | Revised: 8 June 2020 | Accepted: 22 June 2020

DOI: 10.1002/hec.4131

HEALTH ECONOMICS LETTER

Health Economics WILEY

What is it going to be, TTO or SG? A direct test of the validity of health state valuation

Stefan A. Lipman | **Werner B. F. Brouwer** | **Arthur E. Attema**

Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

Correspondence
Stefan A. Lipman, Erasmus School of Health Policy & Management, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands.
Email: lipman@eshpm.eur.nl

Abstract

Standard gamble (SG) typically yields higher health state valuations than time trade-off (TTO), which may be caused by biases affecting both methods. It has been suggested that TTO yields more accurate health state valuations, because TTO is subject to both upward and downward biases that may cancel out. Verifying this claim, however, would require a golden standard to test validity against. In this study, we attempted to provide a first direct test of the validity of health state valuation. A total of 119 students completed five TTO and SG tasks. Afterwards, their health state valuations elicited with TTO and SG were shown to them in an interactive graph. Respondents were asked to indicate which of the methods represented their valuation of a health state best. They could also adjust their valuation. Overall, we found that respondents indicated that TTO valuations better reflected health state valuations, a result that was more pronounced for more severe health states. When offered the opportunity, on average, respondents adjusted health state valuations downwards. These findings may have implications for future work on (bias correction in) health state valuations.

KEYWORDS

feedback module, health state valuation, QALY, standard gamble, time trade-off

1 | INTRODUCTION

Time trade-off (TTO) and standard gamble (SG) are two popular health state valuation methods (Drummond, Sculpher, Claxton, Stoddart, & Torrance, 2015). Both methods enable the estimation of weights representing utility of health status, used for calculating quality-adjusted life years (QALYs). Despite their shared purpose, the operationalization of TTO and SG is different. Perhaps unsurprisingly, so are the health utility weights elicited with these methods (henceforth referred to as QALY weights). Typically, QALY weights elicited with TTO (henceforth TTO weights) are lower than those elicited with SG (henceforth SG weights), which raises questions about which method yields the more appropriate QALY weights (see Bleichrodt & Johannesson, 1997; Lipman, Brouwer, & Attema, 2019b; Torrance, 1976).

Both methods involve direct choices between two options, of which one is living in some health state Q for Y years. In TTO, respondents are offered an alternative: to live in perfect health (described as a state without health

problems) for a shorter time, that is, X years ($X < Y$). Respondents are asked to indicate duration X such that they are indifferent between both options. In practice, this indifference is evaluated by normalizing the utility of perfect health to 1 and that of being dead to 0 and assuming that the utility of duration is linear (Torrance, 1987). Under these strict assumptions, the obtained indifference reveals the TTO weight of state Q, which is obtained by X/Y. SG offers a different alternative to living in health state Q for Y years: a lottery that results in perfect health for Y years (with probability p), or immediate death (with probability $1 - p$). Respondents are asked to indicate probability p such that they are indifferent between both options. In practice, it is often assumed that respondents handle the risky lottery as modeled in expected utility theory. Under this strict assumption, probability p in the obtained indifference reveals the SG weight of state Q.

Bleichrodt (2002) proposed that the differences between TTO and SG weights could be explained by violations of the strict assumptions underlying the methods. SG responses are expected to be biased upwards (by probability weighting and loss aversion), whereas TTO responses are expected to be biased both upwards (by loss aversion) and downwards (by scale compatibility and utility curvature). Hence, Bleichrodt (2002) argued that (i) the upward and downward bias in TTO might cancel out (to some extent) and (ii) the difference between TTO and SG would diminish when these strict assumptions are dropped (e.g., as in prospect theory). Whereas empirical evidence suggests that under less restrictive assumptions, the differences between SG and TTO indeed diminish (Lipman et al., 2019b; Van Osch, Wakker, Van Den Hout, & Stiggelbout, 2004), the first claim by Bleichrodt (2002) is more difficult to address. If bias (partly) cancels out, this implies that that TTO weights are likely to be a better approximation of health state valuation. Testing the validity of QALY weights, however, would require a golden standard for 'true' health state preferences, to compare weights elicited with different methods to. It is safe to say that even the degree to which true preferences exist, and can be measured, is controversial (e.g., Braga & Starmer, 2005), let alone with what method they could be derived.

Instead, we propose a simple, direct test of validity of health state valuation: the opinion of the respondents whose preferences should actually be reflected. After eliciting QALY weights with TTO and SG, we provide respondents with the valuations derived from their responses to reflect on their validity. To our knowledge, our study is the first to ask direct feedback about QALY weights during health state valuation.

2 | METHODS

A total of 119 business administration students took part in this experiment and were rewarded course credit for participation. The sample consisted of 44 (37%) males and 75 (63%) females, with a mean age of 20 (SD = 0.99). The experiment took place in a university-based computer lab, in experimenter-led sessions of 30 min, run with up to four students. The experiment was programmed in Shiny (code available from the first author upon request). In the first and second parts of this experiment, after completing a practice task for a health state described as 'chronic back pain', subjects completed a block of TTO (SG) tasks for the following EQ-5D-5L health states: 21211, 31221, 31231, 31341, and 33342 (see Table 1). Tasks (i.e., TTO and SG) and health states were presented in randomized order between subjects. In the third part (i.e., the 'validation'), subjects were presented with the implications of their responses in the first and second parts (henceforth implied QALY weights) and asked to validate them. After completing the experiment, subjects filled out a paper-and-pencil questionnaire measuring age, sex, and how difficult they felt the tasks were on a scale from 1 (*not difficult at all*) to 10 (*very difficult*).[1]

2.1 | TTO and SG tasks

Both methods were operationalized using the common 10-year duration (as is usual in valuation studies, see Oppe, Devlin, Van Hout, Krabbe, & De Charro, 2014), that is, the period in the impaired health state was 10 years ($Y = 10$). TTO and SG indifferences were obtained through a bisection process with five choices. These five choices produced an indifference point, which subjects could confirm or change with a slider. These slider values were used to calculate the SG and TTO weights (as highlighted in the Section 1). An overview of the instructions and supporting graphs used can be found in the supporting information.

[1]We found no differences in task difficulty across TTO, SG, and validation (paired Wilcoxon tests, p values > 0.48).

TABLE 1 Health states used in this experiment including tariff elicited from Dutch value set for EQ-5D-5L (Versteegh et al., 2016)

Health state	Q1: 21211	Q2: 31221	Q3: 31231	Q4: 31341	Q5: 33342
Dutch tariff	0.88	0.79	0.76	0.47	0.34
You have ... problems with walking	Slight	Moderate	Moderate	Moderate	Moderate
You have ... problems with washing and dressing yourself	No	No	No	No	Slight
You have ... problems with washing and dressing yourself	Slight	Slight	Slight	Moderate	Moderate
... pain or discomfort	No	Slight	Moderate	Severe	Severe
... anxious or depressed	No	No	No	No	No

2.2 | Validation task

For the validation task, subjects were explained the purpose of health state valuation and QALYs (see supporting information). These instructions were based on earlier work in which QALYs were successfully explained in an experimental setting (Bleichrodt, Doctor, & Stolk, 2005). Afterwards, for each health state, the implied QALY weights for TTO and SG (labeled Options A and B and in random order) were visually represented on the QALY scale. Respondents were asked to indicate which value best represented the value of the health state (even if they were the same). Next, respondents could further adjust the chosen QALY weight (henceforth confirmed QALY weights), if they felt that that would improve the health state valuation. To practice, respondents again first completed this validation task for chronic back pain.

3 | RESULTS

For each health state, TTO, SG, and confirmed QALY weights were not distributed normally (Shapiro–Wilk tests, all p values < 0.02). As such, we will apply non-parametric tests and compare median rather than mean QALY weights. Furthermore, the health states used in this experiment allow tests of logical consistency. Each consecutive health state had the same or more problems on each dimension, such that it would be expected that respondents for example preferred state 31221 (Q2) to 31231 (Q3). Only 13% of our sample showed such consistency throughout the whole experiment. The analyses reported below were repeated excluding inconsistent respondents (using a variety of exclusion criteria) and with statistical tests on the mean QALY weights. The main conclusions were unaffected (as shown in the supporting information).

3.1 | TTO and SG tasks

Figure 1 shows that health states received monotonically decreasing QALY weights for both TTO and SG (paired Wilcoxon tests, all p values < 0.001). Interquartile ranges for TTO (SG) were between 0.21 and 0.40 (0.23–0.37) for all health states. The weights were only significantly different from the Dutch EQ-5D tariffs for Q4 and Q5 (Wilcoxon tests,

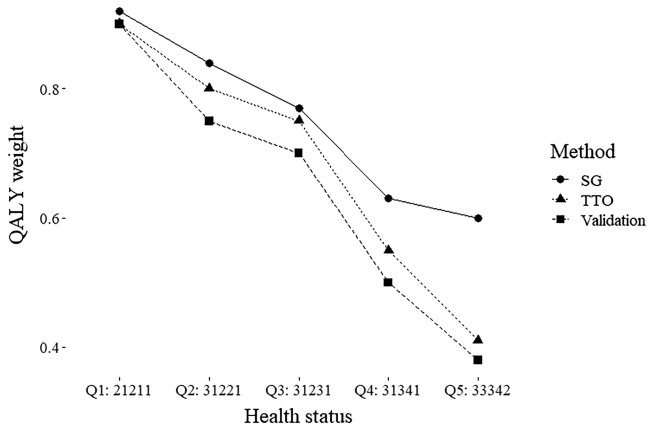

FIGURE 1 Median QALY weights elicited by TTO and SG, including the final QALY weights confirmed after validation. QALY, quality-adjusted life year; SG, standard gamble; TTO, time trade-off

p values < 0.002). For all health states, TTO weights were significantly lower than SG weights (paired Wilcoxon tests, all p values < 0.02). A within-subject comparison showed that for all health states, SG weights were most likely to be higher (55%–66%). Sometimes, TTO values were higher (22%–34%) or both methods yielded equal values (6%–13%). Chi-squared analyses showed that these counts were significantly different (all p values < 0.001), that is, the finding of higher SG values was also confirmed within subjects.

3.2 | Validation task: Preferences for implied QALY weights

Table 2 shows results for the validation task. First, we analyzed which method individuals were more likely to indicate to better represent QALY weights. Table 2 shows the number of individuals who preferred the implied TTO or SG weights per health state. Overall, it can be concluded that for each health state (if a difference existed between TTO and SG), the majority of respondents indicated that TTO weights represented the value of that health state best. Compiled for all health states, this finding was significant (chi-squared test, $p < 0.001$). Analyses per health state showed that it was significant only for the two most severe health states. We also studied preferences for implied QALY weights within subjects, showing that only 18% (5%) consistently preferred TTO (SG) values for all health states. Table 2 also shows

TABLE 2 Overall respondent preferences for implied QALY weights and direction of change for confirmed QALY weights (n denotes the number of respondents, ↑ and ↓ denote upwards and downwards change) and divided by initial ordering of TTO and SG weights

	Overall (n = 119)					
	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Total**
Preferred TTO (n)	63	67	62	78[a]	79[a]	349[a]
Adjusted ↑ (n)	15	12	17	20	16	80
Adjusted ↓ (n)	15	19	21	27	37[b]	119[b]
Preferred SG (n)	56	52	57	41	40	246
Adjusted ↑ (n)	12	8	9	10	8	47
Adjusted ↓ (n)	17	20[b]	22[b]	12	16	87[b]
	Implied TTO weight > SG weight					
	Q1 (n = 40)	Q2 (n = 39)	Q3 (n = 40)	Q4 (n = 38)	Q5 (n = 26)	Total
Preferred TTO (n)	23	14	13	10	8	68
Adjusted ↑ (n)	5	2	6	1	1	15
Adjusted ↓ (n)	5	4	2	5	3	19
Preferred SG (n)	17	25	27[a]	28[a]	18[a]	115[a]
Adjusted ↑ (n)	5	4	6	7	4	26
Adjusted ↓ (n)	6	11	12	12	8	49 [b]
	Implied TTO weight < SG weight					
	Q1 (n = 65)	Q2 (n = 68)	Q3 (n = 72)	Q4 (n = 68)	Q5 (n = 78)	Total
Preferred TTO (n)	33	46[a]	47[a]	59[a]	62[a]	247[a]
Adjusted ↑ (n)	9	8	10	19	12	58
Adjusted ↓ (n)	8	15	19	19	31[b]	92[b]
Preferred SG (n)	32	22	25	9	16	115
Adjusted ↑ (n)	6	4	2	3	4	19
Adjusted ↓ (n)	10	8	8	0	6	32
	Implied TTO weight = SG weight					
	Q1 (n = 14)	Q2 (n = 12)	Q3 (n = 7)	Q4 (n = 13)	Q5 (n = 15)	Total
Preferred TTO (n)	7	7	2	9	9	34
Adjusted ↑ (n)	1	2	1	0	3	7
Adjusted ↓ (n)	2	0	0	3	3	8
Preferred SG (n)	7	5	5	4	6	27
Adjusted ↑ (n)	1	0	1	0	0	2
Adjusted ↓ (n)	1	1	2	0	2	6

Abbreviations: QALY, quality-adjusted life year; SG, standard gamble; TTO, time trade-off.

[a]Signifies that the proportion preferring this method's implied QALY weight (i.e., TTO or SG) is significantly higher than the other (chi-squared test, $p < 0.05$).

[b]Signifies that (of those adjusting) a larger proportion adjusted QALY weights in this direction.

preferences for implied QALY weights and adjustments split for which of the two weights was initially higher. Combined for all health states, these results show that if SG was higher than TTO, respondents were more likely to pick TTO (chi-squared test, $p < 0.001$). The opposite also holds, that is, if TTO was higher than SG, SG weights were more likely to be preferred (chi-squared test, $p < 0.001$). When analyzing separately for each health state, this holds for Q2 (only when TTO weights were smaller than SG weights) to Q5 (chi-squared tests, p values < 0.05). If both weights were the same, no differences were observed in preferred implied QALY weight (chi-squared test, all p values > 0.16), that is, the forced choice among two identical QALY weights was distributed independently.

3.3 | Validation task: Confirmed QALY weights

Respondents were likely to adjust their implied QALY weights, with 50% to 65% choosing to adjust, depending on the health state. Figure 1 shows that for all health states but Q1, median confirmed QALY weights were lower than TTO and SG weights. Interquartile ranges for confirmed QALY weights were similar to TTO and SG, that is, between 0.16 and 0.39, depending on health state. The difference between confirmed QALY weights and TTO weights was significant for Q4 and Q5 (paired Wilcoxon test, p values < 0.02), whereas for SG weights, such significant differences were observed for all health states except Q1 (paired Wilcoxon test, p values < 0.001). Next, we explored the validity of confirmed QALY weights. We found significantly fewer logical inconsistencies after validation compared with both TTO and SG (paired Wilcoxon test, p values < 0.03). Furthermore, combined for all health states, we find the fewest non-trading responses (i.e., QALY weights of 1) for confirmed QALY weights. This non-trading occurred significantly less compared with SG (chi-squared test, $p < 0.001$), but not compared with TTO (chi-squared test, $p = 0.32$). Finally, we explored which types of changes respondents made. Overall, for both TTO and SG, respondents were more likely to change their elicited QALY weights downwards than upwards (chi-squared tests, p values < 0.001). These patterns were also studied separately depending on which method's implied QALY weight was initially higher. Combined for all health states, these analyses show that respondents who made an adjustment were significantly more likely to adjust the lower of the two implied QALY weights downwards, both for TTO (chi-squared test, $p < 0.006$) and SG (chi-squared test, $p < 0.008$).

4 | DISCUSSION

In this study, we provided the first direct test of the validity of health state valuation, by asking respondents to reflect on their QALY weights elicited with TTO and SG. Whereas the EuroQol group by now applies a feedback module in their standard valuation protocol (Stolk, Ludwig, Rand, Van Hout, & Ramos-Goñi, 2019), their module only allows respondents to reflect on the validity of the *ordering* implied by their responses. This ordinal feedback module led to reduced inconsistencies without strongly affecting QALY weights (Wong, Ramos-Goñi, Cheung, Wong, & Rivero-Arias, 2018). Our study goes beyond this approach as we explain the QALY scale and QALY weights to respondents, and in that context, respondents choose which elicited QALY weight is more valid and adjust it if necessary.

We find that, according to respondents themselves, TTO weights are a better reflection of the value of a health state. This finding is in accordance with predictions by Bleichrodt (2002), who argued that upward and downward biases in this method may cancel out. Nonetheless, this canceling out seemingly is not perfect, because respondents often adjusted their implied QALY weights. On average, the direction of this change was downwards, yielding lower confirmed QALY weights than implied by TTO. This may suggest that, as for example was found in Lipman et al. (2019b), the net effect of bias in TTO remains upwards. In this study, the magnitude of this remaining bias appears to increase with severity, with confirmed QALY weights being significantly lower than TTO weights for the two most severe states only. Perhaps, directly correcting biases (Lipman et al., 2019b; Lipman, Brouwer, & Attema, 2019a) could provide a further step towards valid QALY weights, rather than hoping biases in different directions would cancel out.

A few limitations deserve noting. First, we used a non-representative sample consisting of business administration students, which means that our respondents were young, highly educated, with low income, and predominantly female. Seeing as earlier work has suggested that such demographics affect QALY weights (Devlin, Tsuchiya, Buckingham, & Tilling, 2011; Dolan, Gudex, Kind, & Williams, 1996), we encourage future work to apply our approach in (larger) general public samples to test the generalizability of our findings (for instance, in the context of EQ-5D valuation). Although in our student sample the validation task was not considered more complex than TTO and SG, the

description of QALYs used in the validation task might need to be modified (e.g., by using avatar-based explanations, see Lancsar et al., 2020). Second, we used our approach only in the context of health states better than dead, which might be remedied in future work using methods suitable for valuing health states worse than dead. Finally, an alternative explanation of our findings, which also reflects a fundamental difficulty in directly presenting and validating QALY weights, would be that our validation task has some resemblance to a visual analogue scale (VAS). It is well known that VAS may suffer from other biases than TTO and SG do, and QALY weights elicited with VAS are generally lower than those elicited with TTO and SG (Bleichrodt & Johannesson, 1997; Robinson, Dolan, & Williams, 1997; Robinson, Loomes, & Jones-Lee, 2001). As any method of presenting QALY weights might suffer from biases, future work could, for example, test if alternative graphical or textual presentations lead to different conclusions.

To conclude, it appears that, as Bleichrodt (2002) suggested, on average, TTO better reflects individuals' preferences for health states, perhaps as a result of biases canceling out. However, the substantial proportion of individuals who adjusted their QALY weights when given the opportunity suggests that the quest to increase validity of methods in health state valuation methods has not yet ended.

CONFLICT OF INTEREST

None.

ORCID

Stefan A. Lipman https://orcid.org/0000-0002-9507-0612
Arthur E. Attema https://orcid.org/0000-0003-3607-6579

REFERENCES

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, *11*, 447–456. https://doi.org/10.1002/hec.688

Bleichrodt, H., Doctor, J., & Stolk, E. (2005). A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. *Journal of Health Economics*, *24*, 655–678. https://doi.org/10.1016/j.jhealeco.2004.10.001

Bleichrodt, H., & Johannesson, M. (1997). Standard gamble, time trade-off and rating scale: Experimental results on the ranking properties of QALYs. *Journal of Health Economics*, *16*, 155–175. https://doi.org/10.1016/S0167-6296(96)00509-7

Braga, J., & Starmer, C. (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics*, *32*, 55–89. https://doi.org/10.1007/s10640-005-6028-0

Devlin, N. J., Tsuchiya, A., Buckingham, K., & Tilling, C. (2011). A uniform time trade off method for states better and worse than dead: Feasibility study of the 'lead time' approach. *Health Economics*, *20*, 348–361. https://doi.org/10.1002/hec.1596

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996). The time trade-off method: Results from a general population study. *Health Economics*, *5*, 141–154. https://doi.org/10.1002/(SICI)1099-1050(199603)5:2<141::AID-HEC189>3.0.CO;2-N

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*. Oxford university press.

Lancsar, E., Gu, Y., Gyrd-Hansen, D., Butler, J., Ratcliffe, J., Bulfone, L., & Donaldson, C. (2020). The relative value of different QALY types. *Journal of Health Economics*, 102303. https://doi.org/10.1016/j.jhealeco.2020.102303

Lipman, S. A., Brouwer, W. B. F., & Attema, A. E. (2019a). The corrective approach: Policy implications of recent developments in QALY measurement based on prospect theory. *Value in Health*, *22*, 816–821. https://doi.org/10.1016/j.jval.2019.01.013

Lipman, S. A., Brouwer, W. B. F., & Attema, A. E. (2019b). QALYs without bias? Non-parametric correction of time trade-off and standard gamble weights based on prospect theory. *Health Economics*, *28*, 843–854. https://doi.org/10.1002/hec.3895

Oppe, M., Devlin, N. J., Van Hout, B., Krabbe, P. F., & De Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*, *17*, 445–453. https://doi.org/10.1016/j.jval.2014.04.002

Robinson, A., Dolan, P., & Williams, A. (1997). Valuing health status using VAS and TTO: What lies behind the numbers? *Social Science & Medicine*, *45*, 1289–1297. https://doi.org/10.1016/S0277-9536(97)00057-9

Robinson, A., Loomes, G., & Jones-Lee, M. (2001). Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making*, *21*, 17–27. https://doi.org/10.1177/0272989X0102100103

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*, 23–30. https://doi.org/10.1016/j.jval.2018.05.010

Torrance, G. W. (1976). Toward a utility theory foundation for health status index models. *Health Services Research*, *11*, 349–369.

Torrance, G. W. (1987). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases*, *40*, 593–600. https://doi.org/10.1016/0021-9681(87)90019-1

Van Osch, S. M., Wakker, P. P., Van Den Hout, W. B., & Stiggelbout, A. M. (2004). Correcting biases in standard gamble and time tradeoff utilities. *Medical Decision Making*, *24*, 511–517. https://doi.org/10.1177/0272989X04268955

Versteegh, M. M., Vermeulen, K. M., Evers, S. M., De Wit, G. A., Prenger, R., & Stolk, E. A. (2016). Dutch tariff for the five-level version of EQ-5D. *Value in Health*, *19*, 343–352. https://doi.org/10.1016/j.jval.2016.01.003

Wong, E. L., Ramos-Goñi, J. M., Cheung, A. W., Wong, A. Y., & Rivero-Arias, O. (2018). Assessing the use of a feedback module to model EQ-5D-5L health states values in Hong Kong. *The Patient-Patient-Centered Outcomes Research*, *11*, 235–247. https://doi.org/10.1007/s40271-017-0278-0

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lipman SA, Brouwer WBF, Attema AE. What is it going to be, TTO or SG? A direct test of the validity of health state valuation. *Health Economics*. 2020;29:1475–1481. https://doi.org/10.1002/hec.4131