WILEY **Health Economics**

QALYs without bias? Nonparametric correction of time trade-off and standard gamble weights based on prospect theory

Stefan A. Lipman [iD] | Werner B.F. Brouwer | Arthur E. Attema [iD]

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

Correspondence
Erasmus School of Health Policy & Management (ESHPM), Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands.
Email: lipman@eshpm.eur.nl

Abstract

Common health state valuation methodologies, such as standard gamble (SG) and time trade-off (TTO), typically produce different weights for identical health states. We attempt to alleviate these differences by correcting the confounding influences modeled in prospect theory: loss aversion and probability weighting. Furthermore, we correct for nonlinear utility of life duration. In contrast to earlier attempts at correcting TTO and SG weights, we measure and correct all these tenets simultaneously, using newly developed nonparametric methodology. These corrections were applied to three less-than-perfect health states, measured with TTO and SG. We found considerable loss aversion and probability weighting for both gains and losses in life years, and we observe concave utility for gains and convex utility for losses in life years. After correction, the initially significant differences in weights between TTO and SG disappeared for all health states. Our findings suggest new opportunities to account for bias in health state valuations but also the need for further validation of resulting weights.

KEYWORDS
health state valuation, loss aversion, prospect theory, standard gamble, time trade-off

JEL CLASSIFICATION
B41; D03; D81; I10

1 | INTRODUCTION

In cost-utility analyses (CUAs), incremental costs of medical technology are compared with incremental health benefits, commonly expressed in quality-adjusted life years (QALYs). These QALYs (Pliskin, Shepard, & Weinstein, 1980) are obtained multiplying prospective life years by weights, sometimes referred to as "utilities." QALY weights represent health-related quality of life, such that 0 represents the subjective weight of the state "dead" and 1 that of full health. Several methods are used to obtain QALY weights, most notably standard gamble (SG) and time trade-off (TTO). Empirical work, however, has demonstrated that QALY weights differ systematically between these two elicitation methods, with SG weights being higher than TTO weights (e.g., Bleichrodt & Johannesson, 1997; Torrance, 1976). As a consequence, QALY weights and, hence, outcomes of economic evaluations may depend on the health state valuation (HSV) method used.

Bleichrodt (2002) proposed that these discrepancies in elicited QALY weights may result from empirically invalid assumptions present in the theoretical frameworks underlying TTO and SG. More specifically, Bleichrodt argued that TTO and SG weights are biased as they are obtained under the assumptions of expected utility (EU) theory, which has been shown to be descriptively invalid for health outcomes (Bleichrodt, Abellan-Perpiñan, Pinto-Prades, & Mendez-Martinez, 2007; Treadwell & Lenert, 1999). Additionally, although discounted QALY models exist (for an overview, see Hansen & Østerdal, 2006), TTO and/or SG weights are commonly derived under the linear QALY model, which assumes linear utility of life duration (and no discounting of future life years). However, many authors have found diminishing marginal utility of life years; that is, life years that occur in the distant future tend to receive less weight than do life years in the nearer future (Abellan-Perpinan, Pinto-Prades, Mendez-Martinez, & Badia-Llach, 2006; Bleichrodt & Pinto, 2005; Wakker & Deneffe, 1996). In order to obtain QALYs without bias, a methodological shift may be required in HSV towards the use of descriptive utility models such as prospect theory (PT).

PT is characterized by four tenets (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). These are (a) reference dependence—utility derived from a good is defined over differences from a reference point (RP), instead of over the overall consumption of that good; (b) loss aversion—the utility function has an inflection point at the RP and is steeper for losses than for gains; (c) diminishing sensitivity—utility is concave for gains and convex for losses, which indicates diminishing sensitivity to outcomes further from the RP; and (d) probability weighting—the decision maker overweighs small probabilities and underweighs large probabilities (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). PT is usually applied to decisions about money but has also been extended to health outcomes (Bleichrodt & Pinto, 2000; Miyamoto & Eraker, 1989). Importantly, as Bleichrodt (2002) proposed, the tenets modeled in PT will likely affect the TTO and SG methods differently, with loss aversion exerting an upward bias on both methods but utility curvature only affecting TTO whereas probability weighting only affects SG.

Given the increased importance of CUA in informing health policy (Drummond, Sculpher, Claxton, Stoddart, & Torrance, 2015), it is imperative to validly determine the weights that are ascribed to the relevant health states. The valuation of these health states, for example, when obtaining tariffs for the commonly used EuroQol (EQ-5D) generic utility classification system (Versteegh et al., 2016), would necessarily occur within a descriptive context (Bleichrodt, Pinto, & Wakker, 2001). This mean that the status quo of applying EU and/or the linear QALY model to derive TTO and SG weights (a) will not capture actual preferences, as these may include, for example, loss aversion, and (b) may lead to different TTO and SG weights according to Bleichrodt (2002).[1] As such, our main motivation is to address the discrepancy between TTO and SG weights by obtaining these QALY weights using derivations based on a descriptively valid but nonnormative theory (PT). We will refer to this process, where TTO and SG weights are obtained while incorporating loss aversion, nonlinear utility, and/or probability weighting into their derivation, as *correction for PT*. If correcting TTO and SG for PT is feasible, it could be used to correct observed responses in HSVs, allowing corrected weights to be used when calculating QALYs to express health benefits in CUAs, as commonly done.

Some studies have attempted to test Bleichrodt's (2002) predictions about PT and correct HSV techniques by assuming PT or adjusting for utility curvature (Attema & Brouwer, 2009; Martin, Glasziou, Simes, & Lumley, 2000; Oliver, 2003; van Osch, Wakker, van den Hout, & Stiggelbout, 2004; Wakker & Stiggelbout, 1995). Yet to date, no study has been able to simultaneously correct both TTO and SG for loss aversion, utility curvature, and probability weighting (see Appendix S1 for an overview of earlier studies on corrections). In this study, we adapted a recently proposed methodology (Abdellaoui, Bleichrodt, L'Haridon, & Van Dolder, 2016) to measure these three deviations without parametric assumptions and elicit TTO and SG weights without assuming EU or the linear QALY model. In other words, we provide the first empirical test of predictions by Bleichrodt (2002) and show how correcting for PT alleviates the discrepancies between TTO and SG.

Our study features several methodological improvements compared with previous attempts at correcting TTO and/or SG weights for PT (see Appendix S1). First, our adaptation of the nonparametric method (Abdellaoui et al., 2016) enables us to determine utility curvature, loss aversion, and probability weighting separately for each individual, without assuming a specific parameter or parametrical form for these functions (as opposed to work by van Osch et al., 2004, Martin et al., 2000, van der Pol & Roux, 2005). We believe this is relevant, as large heterogeneity typically exists for PT elicitations (Pinto-Prades & Abellan-Perpiñan, 2012), warranting an individual measurement approach. Furthermore, applying specific parametric forms within experimental elicitation can confound results (Abdellaoui, 2000), thus

---

[1]These statements hold regardless if one believes EU to be the normative standard (as Kahneman & Tversky, 1979, and Wakker, 2010, do), which would, for example, classify loss aversion as "irrational" or a bias. We will make no such claims and will refer to deviations of EU and the linear QALY model as generating bias in TTO and SG.

allowing considerable bias to remain after correction (Wakker, 2008; Wakker, 2010). Second, we attempt to append the heterogeneity surrounding RPs by providing all subjects with the same RP, which is a hypothetical expected life duration (following the successful procedure described in Attema, Brouwer, & L'Haridon, 2013). This is important, because even though reference dependence appears to be the most central tenet of PT, earlier work on the location of the RP suggests that individuals use multiple different health outcomes as RP (Bleichrodt et al., 2001; van Osch et al., 2004; van Osch & Stiggelbout, 2008; van Osch, van den Hout, & Stiggelbout, 2006).

2 | THEORETICAL FRAMEWORK

We describe health outcomes as (β, t), where β represents health status and t indicates the age at which the health profile ends (e.g., living with chronic back pain until 70). Throughout, subscripts (e.g., x and y) are used to refer to possible health profiles faced by a single agent, with age of onset (e.g., current age) denoted by t_a. We will often suppress t_a by denoting (β_x, t_x) as (β_x, T_x), with duration defined by $T_x = t_x - t_a \geq 0$. We refer to (β_x, T_x) as chronic health profiles. We let $(\beta_x, T_x)_p(\beta_y, T_y)$ denote the risky prospect that provides health profile (β_x, T_x) with probability p and health profile (β_y, T_y) with probability $1 - p$. Preferences are denoted using the conventional notations $\succ$, $\succeq$, and $\backsim$ to represent strict preference, weak preference, and indifference, respectively. Also, we assume weak-ordered preferences; that is, they are complete, meaning that decision makers have preferences over risky prospects, and transitive (if $x \succeq y$ and $y \succeq z$, then $x \succeq z$). Health profiles (β_x, T_x) starting and ending at t_a (so that $t_a = t_x$) will thus have $T_x = 0$ (i.e., they equal immediate death), and, for brevity, we will denote such profiles of the form $(\beta_x, 0)$ as D, for any β_x. As in Miyamoto, Wakker, Bleichrodt, and Peters (1998), we assume indifference between all profiles denoted D for any β. Finally, we assume monotonicity for duration, that is, $(\beta_x, T_x) \succ (\beta_x, T_y)$ for $T_x > T_y$ and any β_x.

The general QALY model assumes that preferences for health profiles (β_x, T_x) are represented by the general utility function $V(\beta_x, T_x) = U(\beta_x) * L(T_x)$. In this model, $L(T)$ and $U(\beta)$ denote utility functions over life years or health status, respectively. This QALY model, and the preference foundations underlying it, typically relies on EU to some extent (for axiomatizations, see Miyamoto & Eraker, 1989, Miyamoto & Eraker, 1988). To derive corrected TTO and SG weights, we will extend this model to incorporate insights from PT under risk. That is, we assume that preferences can be represented by the general QALY model, including the extensions we outline below.

Several preliminaries are required before defining our full model (Equations (1) and 2). We assume that preferences for health profiles are defined relative to an RP, which we denote as (β_r, T_r). Following Wakker (2010), we define this RP as a point of comparison, which may differ during different parts of the analysis. Given that no plausible theory of RP selection is available (Wakker, 2010), we let the RP depend on framing of the decision context. Hence, (β_r, T_r) refers to an expected health profile described in a decision task, which is taken as the neutral point. This health profile has health status β_r, endured for T_r years. Throughout, for brevity, we denote the duration of all other health profiles as deviations from the RP; that is, we denote health profiles (β_x, T_x) as (β_x, T_x^*) with $T_x^* = T_x - T_r$ in β_x. We will restrict our model to health profiles $(\beta_x, T_x^*) \succeq D$ with $\beta_x \succeq \beta_r$ for any T_x^*. In other words, we assume our model holds for a restricted outcome domain including only health profiles weakly preferred to immediate death, where health status remains at β_r or is improved.

Within this outcome domain, we model PT by incorporating sign dependence for life duration, that is, by modifying $L(T)$ in the general QALY model to $L^i(T^*)$. In our model, $L^i(T^*)$ is a standard, real-valued ratio scale utility function with $L^+(T_r) = 0$, which may be different for gain outcomes $(\beta_x, T_x^*$, with $\beta_x \backsim \beta_r$ and $T_x^* \geq 0)$ and loss outcomes $(\beta_x, T_x^*$, with $\beta_x \backsim \beta_r$ and $T_x^* < 0)$. We do not modify $U(\beta)$ in our model, which implies that changes in health status will be evaluated as in the conventional general QALY model. We incorporate loss aversion[2] by taking $L^-(T^*) = \lambda L^i(T^*)$ for $T^* < 0$. Here, λ denotes a loss aversion index, with $\lambda > 1$ ($\lambda = 1$, $\lambda < 1$) indicating loss aversion (loss neutrality, gain seeking). Furthermore, we incorporate nonlinear weighting of probabilities by incorporating probability weighting functions $w^i(p)$, $i = +, -$, for gains and losses respectively, that assign a number to each probability p, with $w^i(0) = 0$ and $w^i(1) = 1$.

We will apply this model to risky prospects with at most two outcomes, that is, binary prospects. Thus, preferences over risky prospects with both gain and loss outcomes, that is, $(\beta_x, T_x^*)_p(\beta_y, T_y^*)$, with $T_x^* \geq 0 > T_y^*$ are evaluated by

---

[2]In our simplified approach, we model PT over life duration by assuming attribute-specific evaluation (as in Bleichrodt et al., 2009). Loss aversion is, thus, defined over life duration, as it is not meaningful on $U(\beta_x)$ when health status is considered a qualitative measure (Bleichrodt and Miyamoto, 2003). This does not affect our analysis, as we only consider improvements in health status.

$$w^+(p)U(\beta_x)L^+\left(T_x^*\right) + w^-(1-p)U\left(\beta_y\right)L^-\left(T_y^*\right),$$ (1)

whereas preferences over risky prospects $\left(\beta_x, T_x^*\right)_p\left(\beta_y, T_y^*\right)$ for either gains or losses are evaluated by

$$w^i(p)U(\beta_x)L^i\left(T_x^*\right) + \left(1 - w^i(p)\right)U\left(\beta_y\right)L^i\left(T_y^*\right), \, i = +, -,$$ (2)

where $i = +$ [$-$] when $T_x^*, T_y^* > [<] 0$, that is, both outcomes are gains or losses. Whenever $w^i(p) = p$, $\lambda = 1$, and no distinction is made between gains and losses (i.e., no reference dependence), our model reduces to the general QALY model.

2.1 | SG and TTO correction for PT

TTO weights are obtained by eliciting duration T_y, which yields indifference between (β_x, T_x) and (FH, T_y), with $T_x > T_y$. SG weights, on the other hand, are obtained from indifferences between a certain outcome (β_x, T_x), and a risky prospect $(FH, T_x)_p(D)$, where p is normally varied until indifference is obtained. Often, TTO and SG weights (i.e., $U(\beta_x)$) are derived under the assumptions of EU and the linear QALY model, which is a special case of the general QALY model with $L(T) = T$, $U(FH) = 1$, and $V(D) = 0$. Under these assumptions, indifferences $(\beta_x, T_x) \sim (FH, T_y)$ and $(\beta_x, T_x) \sim (FH, T_x)_p(D)$ allow derivation of TTO and SG weights for health state β_x by $U(\beta_x) = \dfrac{T_y}{T_x}$ and $U(\beta_x) = p$, respectively.

Our correction for PT involves deriving TTO and SG weights by means of our theoretical model based on PT. The application of our theoretical model requires assumptions about the RP used in TTO and SG. Typically, TTO and SG exercises are framed with the impaired health state (β_x, T_x) as RP. Furthermore, earlier work on SG[3] has suggested that the outcome that remains constant, that is, the time spent with reduced health status (β_x, T_x), usually is taken as RP (Bleichrodt et al., 2001; van Osch et al., 2006). Hence, throughout the paper, we will make the following assumption about the RP for TTO and SG: $(\beta_r, T_r) = (\beta_x, T_x)$.

Under these assumptions, TTO indifferences $(\beta_x, T_x) \sim (FH, T_y)$ allow the following derivation for $U(\beta_x)$:[4]

$$U(\beta_x) = \frac{L^-\left(T_y^*\right) + 1}{(1-\lambda)L^-\left(T_y^*\right) + 1},$$ (3)

whereas SG indifference $(\beta, T_x) \sim (FH, T_x)_p(D)$ allows the following derivation for $U(\beta_x)$ as in Bleichrodt et al. (2001):

$$U(\beta_x) = \frac{w^+(p)}{w^+(p) + \lambda w^-(1-p)}.$$ (4)

2.2 | Parameter elicitation

In order to correct both TTO and SG weights for PT, that is, to be able to compute the outcome of Equations (3) and (4), one needs to elicit the following: (a) $L^i(T^*)$ with T_x^* as RP to allow estimation of $L^-\left(T_y^*\right)$, (b) probability weighting functions $w^i(p)$, $i = +, -$, and (c) a loss aversion coefficient λ, which reflects overweighting of losses with T_x^* as RP. This means that t_x should be kept constant across TTO and SG and the elicitation of $L^i(T^*)$, to ensure that λ refers to the same theoretical construct throughout (i.e., the same kink around the RP, see Section 4.4).

---

[3]No empirical work exists studying the RP for TTO. Here, we assumed that it coincides with that of SG and with how TTO is typically framed. If the time spent in perfect health (i.e., FH, T_y) is taken as RP instead, Equation 3 cannot be applied. This also holds for SG; that is, Equation 4 is only valid if the RP is actually (β_x, T_x).

[4]Equations 3 and (4) apply a scaling of $L^i(T^*)$, where the utility of the lowest outcome is set to -1, for simplicity (i.e., $L^-(T_a) = -1$). For elaborate proofs of Equations 3 and (4) under our theoretical model, see Appendix S2.

3 | METHODS

We report the results of an experiment in which we compare TTO and SG weights derived assuming EU and the linear QALY model to QALY weights corrected for PT (i.e., by Equations (3) and 4). In this experiment, PT parameters were elicited using methodology based on the work by Abdellaoui et al. (2016). To reduce the influence of order effects and test for consistency, multiple counterbalancing procedures were conducted between participants and consistency checks were in place (see Appendix S3). The experiment was computerized in Matlab. Subjects were 99 students of the Rotterdam School of Management (58 female) who were rewarded course credits. Experimental sessions lasted for approximately 55 min and were run on computers in sessions of four subjects sitting adjacently in separate cubicles. An instructor was present at all times to answer questions.

3.1 | TTO and SG weight elicitation

We elicited TTO and SG weights for a total of four health states (one practice state) from the EQ-5D-5L (five level) descriptive system (Herdman et al., 2011). These health states reflected an array of mildly aversive health states, in order to avoid health states that could be considered worse than death (Dolan, 1997). The following health states were used: 22222 (practice, β_p), $\beta_1 = 21211$, $\beta_2 = 31221$, and $\beta_3 = 32341$. We applied a bisection choice-based elicitation procedure with four consecutive choices, as choice-based procedures produce more consistent measurements than matching (Noussair, Robin, & Ruffieux, 2004). Subjects were asked to imagine having lived until age 50 in perfect health after which they contracted a disease that would affect their quality of life for their remaining life expectancy of 20 years. TTO and SG were completed for these remaining 20 years (i.e., $t_a = 50$). In both cases, the maximum expected age of death was 70 years; that is, subjects made decisions with regard to the quality of life for age 50 to 70 (followed by death), which ensured that t_x was constant for both TTO and SG.

3.2 | Nonparametric method

We adapted Abdellaoui et al.'s (2016) nonparametric methodology to measure PT under risk in the health domain. In order to elicit $L^i(T^*)$ with the same t_x as RP as in TTO and SG, we instructed subjects to take living from current age until 70 in perfect health as RP, that is, $(\beta_r, T_r) = (FH, 70 - t_a)$. Elicitation consisted of four stages (an elaborate description of the method and instructions can be found in Appendices S1, S4, and S5). The first stage connected utility for gains ($L^+(T^*)$) to the utility for losses ($L^-(T^*)$). The second and third stages employed the trade-off method of Wakker and Deneffe (1996) to measure a standard sequence of utility for gains and utility for losses, respectively. The fourth stage measured probability weighting, separately for gains and losses; that is, $w^+(p)$ and $w^-(p)$. Our methodology thus makes it possible to completely elucidate PT's tenets in the health domain, without imposing parametric assumptions on $L^i(T^*)$ and $w^i(p)$. Each of the four stages had slightly different instructions (see Appendix S5), providing the context for the trade-offs that subjects were required to make. Subjects had to choose between two medicines that could amend their situation but would not affect their life expectancy, which remained constant at perfect health. All indifferences were elicited using a bisection choice-based procedure with a slider (following Abdellaoui et al., 2016) where subjects first performed three binary choices. This procedure zoomed in to the point at which subjects would become indifferent but still allowed subjects to specify the final value and adjust accordingly. To allow estimation of $L^-\left(T_y^*\right)$ in Equation (3) regardless of the amount of years given up in TTO, subjects' standard sequence continued to at least 20 years above and below t_x (i.e., living until 70), to avoid extrapolation beyond the measured curve[5].

3.3 | Analyses of curvature for $L^i(T)$

We used two methods to investigate the curvature of $L^i(T^*)$, that is, utility curvature: a nonparametric method and a parametric method (similar to Abdellaoui et al., 2016). For these analyses of utility curvature, we normalized all durations by dividing through subjects' highest absolute elicited duration for gains and losses, respectively (T_{kG}^* or $-T_{kL}^*$).

---

[5]After 25 steps, the standard sequence elicitation was terminated to avoid overburdening our subjects. When necessary, $L^-\left(T_y^*\right)$ was obtained by extrapolation.

This resulted in T^* being in the range $[-1, 1]$. Next, we calculated the area under the curve (AUC) of $L^i(T^*)$ separately for both domains, by setting $L^+(T^*_{kG}) = 1$ and $L^-(T^*_{kL}) = -1$. If utility of life duration is linear, the area under this normalized curve equals one half. Utility for gains in life duration is convex (concave) if the AUC is smaller (larger) than one half, whereas for losses, the opposite direction holds (convex > ½, concave < ½). This method of analyzing utility curvature is nonparametric. We also analyzed $L^i(T^*)$ parametrically by employing the most commonly used power utility family using nonlinear least squares, using the same normalizations. For this family, $L^+(T^*) = (T^*)^\alpha$ and $L^-(T^*) = -((T^*)^\alpha)$ with $\alpha > 0$. For gains [losses], $\alpha > 1$ corresponds to convex [concave] utility, $\alpha = 1$ corresponds to linear utility, and $\alpha < 1$ corresponds to concave [convex] utility.

3.4 | Analyses of loss aversion

Several definitions of loss aversion exist, with λ being interpreted in various manners (see Köbberling & Wakker, 2005). Köbberling and Wakker (2005) defined loss aversion (λ) as the kink of utility at the RP. That is, they define loss aversion as $U'_\uparrow(0)/U'_\downarrow(0)$, with $U'_\uparrow(0)$ representing the left derivative and $U'_\downarrow(0)$ the right derivative of U at the RP. Hence, we computed each subject's coefficient of loss aversion (λ) over the first steps in their standard sequence for gains and losses, denoted as x_1^+ and x_1^-. Loss aversion is then defined as the ratio of $L^-(x_1^-)/x_1^-$ over $L^+(x_1^+)/x_1^+$, which is equal to $x_1^+/-x_1^-$ (Abdellaoui et al., 2016). A subject was classified as loss averse if $x_1^+/-x_1^- > 1$, loss neutral if $x_1^+/-x_1^- = 1$, and gain seeking if $x_1^+/-x_1^- < 1$ (as in Wakker, 2010).

3.5 | Probability weighting

We used certainty equivalences using varying probabilities to elicit the weighting functions, similar to Attema, Bleichrodt, and L'haridon (2018). In particular, we used linear interpolation to obtain a $w^+(p)$ and $w^-(p)$, using $p = 0.1, 0.3, 0.5, 0.7, 0.9$. Furthermore, we used Tversky and Kahneman's one-parameter inverse S-shaped probability weighting function $w^i(p) = p^\gamma/(p^\gamma + (1-p)^\gamma)^{1/\gamma}$ with $i = +, -$, estimated by nonlinear least squares. The γ-parameter controls for the shape of the probability weighting function. If $\gamma = 1$, there is no probability transformation and $w^i(p) = p$. However, if $\gamma < 1$, decision makers underweight large probabilities and overweight small probabilities. This corresponds to the commonly found inverse S-shaped weighting function. If $\gamma > 1$, the opposite pattern holds, corresponding to an S-shaped weighting function.

4 | RESULTS

Two subjects expressed unwillingness to trade off any life years, which caused the experiment to fail. These subjects were removed from further analyses. As can be seen in Appendix S3, we included several repetitions to test for consistency. At the aggregate level, we observed significant differences between the consistency indifference value and the value for x_2^i (i.e., the second step) in the standard sequence elicitation for both gains and losses (paired t tests: $ps < .01$). Furthermore, we found a difference for the consistency checks in the probability sequence for gains (paired t test: $ps = .007$), but not for losses (paired t tests: $ps = .62$). Correlations between consistency checks and original values were high, suggesting strong association between these values (Kendall's τs > 0.51, $ps < .003$).

Twenty-nine subjects violated monotonicity for health states, which indicates that they valued at least one health state, which was better or equal on each dimension lower than their dominated counterpart (e.g., 21211 vs. 31221). As we consider that it is plausible that all subjects prefer more health to less, we reran the full analyses excluding these subjects and found no differences in the main results. Hence, we report the results for the full sample ($n = 97$).

4.1 | Curvature of $L^+(T)$ and $L^-(T)$

We observed median AUC for gains equal to 0.555, and for losses, this nonparametric analysis produced a median AUC of 0.561, which were both significantly different from 0.5 (Wilcoxon signed ranks tests: $ps < .001$). After parametrically fitting a power function to the data, we found a median α of 0.787 for gains and 0.757 for losses (significantly smaller

than 1, Wilcoxon signed ranks tests: $ps < .001$). Thus, both parametric and nonparametric results demonstrated $L^+(T^*)$ to be concave and $L^-(T^*)$ to be convex.

Table 1 shows the classification of subjects' curvature for gains ($L^+(T^*)$) and losses ($L^-(T^*)$) at the individual level, both parametrically and nonparametrically. The most common pattern was concave curvature for $L^+(T^*)$ and convex curvature for $L^-(T^*)$ as was found in an earlier implementation of this method (Attema et al., 2018). This conclusion holds for both nonparametric (53%) and parametric (53%) results.

4.2 | Loss aversion

Utilizing Köbberling and Wakker's (2005) definition, we found a median loss aversion index of $\lambda = 2$ (interquartile range: 1.00–3.52). Thus, we found considerable loss aversion at the aggregate level, with the median being significantly higher than 1 (Wilcoxon test: $p < .001$). At the individual level, the majority of subjects demonstrated loss aversion, with 72% ($n = 70$) classifying as loss averse, and 15% ($n = 15$) and 13% ($n = 12$) classifying as loss neutral or gain seeking, respectively.

4.3 | Probability weighting ($w^i(p)$)

Figure 1 shows the median decision weights assigned to $p = 0.1, 0.3, 0.5, 0.7, 0.9$. As can been seen from the plots, we observe inverse S-shaped probability weighting for both gains and losses, with more pronounced overweighting of small probabilities for losses. Using Tversky and Kahneman's one-parameter function, we found a median $\gamma = 0.92$ for gains and a median $\gamma = 0.84$ for losses (both significantly lower than 1, Wilcoxon tests: $ps < .04$). Both analyses demonstrated that the typical inverse S-shaped probability transformation was the most prevalent in our data, for both gains and losses. Moving to the individual level, for gains, we found $\gamma < 1$ for 56 subjects (58%) and $\gamma > 1$ for 41 subjects (42%). For losses, we found more pronounced inverse S-shaped probability weighting, with 71 (73%) and 26 (27%), respectively.

TABLE 1 Classification for curvature of $L^+(T^*)$ and $L^-(T^*)$ at the individual level

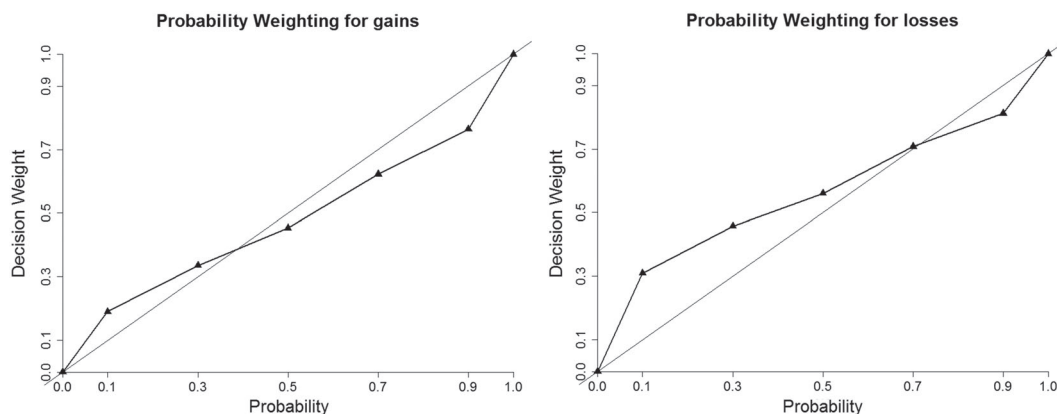| Gains $L^+(T^*)$	Losses—$L^-(T^*)$			
	Concave	Convex	Linear	Total
		Nonparametric		
Concave	19	51	0	70
Convex	7	17	1	25
Linear	0	1	1	2
		Parametric		
Concave	19	51	0	70
Convex	6	18	1	25
Linear	0	1	1	2

FIGURE 1 Probability weighting functions for gains ($w^+(p)$) and losses ($w^-(p)$)

4.4 | Health state correction

Table 2 shows QALY weights for all health states elicited using TTO and SG, where uncorrected refers to weights elicited assuming EU and linear QALYs, whereas corrected weights are elicited by means of Equations (3) and (4). To test the sensitivity of our results to linear interpolation, we also corrected TTO and SG weights by using power utility to estimate $L^-\left(T_y^*\right)$ and the Kahneman and Tversky probability weighting function to estimate $w^+(p)$ and $w^-(1-p)$; these are indicated by "Parametric Corrections" in Table 2. An initial difference in TTO and SG weights existed (paired t tests, all $ps < .001$), with SG weights being higher than TTO for all β_x. Our results show that the corrected weights were lower than the uncorrected weights for TTO and SG (paired t tests: all $ps < .01$). The initially significant difference between the uncorrected weights only disappeared for all β after applying nonparametric corrections (paired t tests: all $ps > .09$). The parametric corrections left significant and substantial differences between TTO and SG weights.

Finally, we performed four isolated corrections. For the sake of brevity, we only report the results of the nonparametric corrections (see the Supporting Information for results of these analyses for parametric corrections). First, we corrected TTO for utility curvature only, with $\lambda = 1$. Second, TTO weights were corrected for loss aversion only, with linear utility (i.e., $L^i(T^*) = T^*$). Third, we corrected SG for probability weighting only, with $\lambda = 1$. Finally, SG weights were corrected for loss aversion only, with $w^i(p) = p$. This allows us to demonstrate the influence of each correction in isolation. Table 3 shows that correcting for loss aversion had a stronger downward influence on TTO weights than correcting for curvature of $L^i(T^*)$, and both correcting for probability weighting and correcting for loss aversion had a substantial negative influence on SG weights.

5 | DISCUSSION

This paper provides the first empirical test of Bleichrodt's (2002) predictions about PT, demonstrating that it may be possible to correct the weights typically used in HSV, that is, to reduce bias in TTO and SG.

TABLE 2 Overview of mean weights [standard deviation] for health states β_{1-3} for TTO and SG including differences between methodologies under multiple corrections

Correction	Health state	TTO weight	*SD*	SG weight	*SD*	Difference
Uncorrected	β_1: 21211	0.665	[0.268]	0.75	[0.25]	−0.085*
	β_2: 31221	0.605	[0.259]	0.706	[0.261]	−0.101*
	β_3: 32341	0.39	[0.259]	0.518	[0.276]	−0.128*
Nonparametric	β_1: 21211	0.492	[0.331]	0.506	[0.295]	−0.014 ns
	β_2: 31221	0.442	[0.313]	0.456	[0.287]	−0.014 ns
	β_3: 32341	0.279	[0.27]	0.319	[0.229]	−0.039 ns
Parametric	β_1: 21211	0.496	[0.325]	0.598	[0.319]	−0.102*
	β_2: 31221	0.449	[0.307]	0.558	[0.322]	−0.109*
	β_3: 32341	0.295	[0.272]	0.387	[0.303]	−0.092*

Abbreviations: SG, standard gamble; TTO, time trade-off.

*Differences were significant at $p < .001$ for paired t tests.

TABLE 3 Isolated effects of corrections for UC, LA, and PW for TTO and SG weights [standard deviation in brackets]

Health state	Uncorrected weight		UC only		LA only		PW only	
TTO: Implication	$\lambda = 1$ and $L^i(T^*) = T^*$		$\lambda = 1$		$L(T^*) = T^*$			
β_1: 21211	0.665	[0.268]	0.611	[0.296]	0.537	[0.311]		
β_2: 31221	0.605	[0.259]	0.558	[0.287]	0.474	[0.3]		
β_3: 32341	0.39	[0.259]	0.364	[0.278]	0.288	[0.259]		
SG: Implication	$\lambda = 1$ and $w^i(p) = p$				$w^i(p) = p$		$\lambda = 1$	
β_1: 21211	0.75	[0.25]			0.63	[0.307]	0.643	[0.246]
β_2: 31221	0.706	[0.261]			0.584	[0.305]	0.597	[0.249]
β_3: 32341	0.518	[0.276]			0.387	[0.278]	0.459	[0.218]

Abbreviations: LA, loss aversion; PW, probability weighting; SG, standard gamble; TTO, time trade-off; UC, utility curvature.

We estimated the full set of PT's parameters in the health domain, in order to obtain more descriptively valid outcomes, which can be used in the QALY model. Our results are consistent with PT (Kahneman & Tversky, 1979): We observe concave utility curvature for gains and convex utility curvature for losses, inverse S-shaped probability weighting, and considerable loss aversion. In general, the estimates of utility curvature for gains in life duration and loss aversion (when applicable) of earlier work are similar to ours (e.g., Attema, Brouwer, & L'Haridon, 2013; Bleichrodt & Pinto, 2000; Bleichrodt & Pinto, 2005), but different results are found for the utility function for losses in life duration. These differences might be explained by methodological differences, which is a hypothesis that could be tested in future work. Furthermore, we replicated the typical finding that SG weights are higher than TTO weights. By means of corrections similar to those proposed by Bleichrodt et al. (2001), we attempted to remove the systematic bias in these weights, by simultaneously accounting for loss aversion, probability weighting, and utility curvature. Consequently, as predicted by Bleichrodt (2002), the weights assigned to both TTO and SG were markedly lower than their uncorrected counterparts. Moreover, they were no longer significantly different.

Although successful attempts at correcting SG and/or TTO weights using parametric methodology are reported in earlier work (Martin et al., 2000; van der Pol & Roux, 2005; van Osch et al., 2004), our parametric corrections were not able to fully account for the discrepancies between these methods. This seemed to be driven by SG weights remaining higher when parametric estimations for probability weighting were used. Given that our nonparametric estimations of probability weighting allowed full flexibility of the weighting function (see Abdellaoui, 2000), these findings suggest that parametric estimations of probability weighting may produce different results.

Our results demonstrate that, considered in isolation, loss aversion had a stronger downward influence on TTO weights than utility curvature, whereas both probability weighting and loss aversion lowered SG weights considerably. Although these findings are generally in line with previous studies, we observed a downward effect of correcting TTO for utility curvature. Probably, this is caused by the convexity found for losses in life years and the framing of our TTO and SG exercises (which both featured losses in life years from the RP in a reduced health state). Future work could shed light on the degree to which this discrepancy may be caused by the nonparametric method or the framing used in our work.

Several limitations of our study need noting. First, several subjects violated monotonicity for the health states used. Although excluding these subjects from the sample did not alter our results, we expect that these errors in decision making are to be attributed to either (a) imprecision of preferences or (b) error propagation, that is, early errors cascading into later stages of the task. Considering the use of only relatively mild health states, for which subjects may have no precise preference ordering in mind, some overlap may occur within our method. Regarding error propagation, it is good to note that during utility elicitation, subjects could rectify errors by adjusting the final indifference value on the slider to any nondominant value in life years, that is, fix their earlier "errors." Testing for error propagation, by performing an error simulation as described by Bleichrodt and Pinto (2000), confirmed that errors did not have a propagating effect on the standard sequence we elicited for gains and losses.[6]

Second, concerns may be raised about the role of the RP in this paper. We find that the observed discrepancies between TTO and SG can be removed by correcting under the assumption that decision makers utilize the guaranteed outcome (β_x, T_x) as RP (which ensures that t_x remains constant). However, earlier work on health-related preferences has suggested that individuals may also use their own current health and life expectancy as RP (van Nooten & Brouwer, 2004; van Nooten, Koolman, & Brouwer, 2009). In our work, we found no evidence of such effects.[7] A related limitation concerns our assumption that subjects use the fixed outcome in both TTO and SG as their RP, which is crucial for our results as our corrections depend on a constant T_r throughout the multiple parts of the experiment. Earlier work, however, demonstrated that SG subjects may also use the time spent in full health as their RP (van Osch & Stiggelbout, 2008). To our knowledge, such work does not exist for TTO methods. Therefore, future work should explore the possibility of correcting under the assumption that subjects use full health as RP, for both TTO and SG.

Finally and perhaps most importantly, the primary goal of the present research was merely to provide the first empirical test of Bleichrodt's (2002) predictions for TTO and SG weights, and our findings should be interpreted in this context. We observed considerable differences to nationally representative findings. For example, the Dutch tariff (Versteegh et al., 2016) for health state β_1 (21211) is 0.876, whereas we elicited a raw TTO weight of 0.665. Our sample,

---

[6]The difference between TTO and SG weights not was not significant in all simulations ($k = 1,000$) for β_1 and β_2, while replicating our results in the majority of simulations for β_3 (over 70%). These simulations suggest that our correction method is quite robust to error propagation.

[7]We tested for associations between subjects' self-reported life expectancy and their estimates for loss aversion, utility curvature, and probability weighting; no such associations were observed for raw and corrected health state weights (all Kendall's τs < 1.52, all ps > .13).

consisting of young, healthy students will have contributed strongly to this initial discrepancy, next to differences in methodology. We also note that after correction, the discrepancy between tariffs and corrected weighs increases. After the nonparametric correction, the QALY value of state β_1 decreases to 0.492. Clearly, this calls for further investigation of the methods used here, also in other (general public) samples, in order to further explore the impact of corrections and further refine the methods used. This future research may also clarify whether our framing may have yielded relatively low weights and how the methods used here can be simplified to be suitable for use in general public samples.

6 | CONCLUSION

With the increasing importance of economic evaluations in health care, the question of how to best estimate health states valuations has become a crucial one. Conventional methodologies, such as TTO and SG, systematically arrived at different valuations of the same health state. PT may offer an explanation for this phenomenon (Bleichrodt, 2002), which was never tested directly. Using the nonparametric method (Abdellaoui et al., 2016), we demonstrated that it may be possible to significantly reduce these biases in HSVs. After correction for loss aversion, probability weighting, and utility curvature, TTO and SG weights for three health states were no longer different. This is an encouraging finding, but at the same time, the resulting low absolute values highlight the need for future research. Notwithstanding these important limitations, our findings do suggest the feasibility and relevance of this approach and may prove to be a first step in the move towards QALYs without bias.

CONFLICTS 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

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, *46*, 1497–1512. https://doi.org/10.1287/mnsc.46.11.1497.12080

Abdellaoui, M., Bleichrodt, H., L'Haridon, O., & Van Dolder, D. (2016). Measuring loss aversion under ambiguity: A method to make prospect theory completely observable. *Journal of Risk and Uncertainty*, *52*, 1–20. https://doi.org/10.1007/s11166-016-9234-y

Abellan-Perpinan, J. M., Pinto-Prades, J. L., Mendez-Martinez, I., & Badia-Llach, X. (2006). Towards a better QALY model. *Health Economics*, *15*, 665–676. https://doi.org/10.1002/hec.1095

Attema, A. E., Bleichrodt, H., & L'haridon, O. (2018). Ambiguity preferences for health. *Health Economics*, *27*(11), 1699–1716.

Attema, A. E., & Brouwer, W. B. (2009). The correction of TTO-scores for utility curvature using a risk-free utility elicitation method. *Journal of Health Economics*, *28*, 234–243. https://doi.org/10.1016/j.jhealeco.2008.10.004

Attema, A. E., Brouwer, W. B., & L'Haridon, O. (2013). Prospect theory in the health domain: A quantitative assessment. *Journal of Health Economics*, *32*, 1057–1065. https://doi.org/10.1016/j.jhealeco.2013.08.006

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., Abellan-Perpiñan, J. M., Pinto-Prades, J. L., & Mendez-Martinez, I. (2007). Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science*, *53*, 469–482. https://doi.org/10.1287/mnsc.1060.0647

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

Bleichrodt, H., & Miyamoto, J. (2003). A characterization of quality-adjusted life-years under cumulative prospect theory. *Mathematics of Operations Research*, *28*(1), 181–193.

Bleichrodt, H., & Pinto, J. L. (2000). A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science*, *46*, 1485–1496. https://doi.org/10.1287/mnsc.46.11.1485.12086

Bleichrodt, H., & Pinto, J. L. (2005). The validity of QALYs under non-expected utility. *The Economic Journal*, *115*, 533–550. https://doi.org/10.1111/j.1468-0297.2005.00999.x

Bleichrodt, H., Pinto, J. L., & Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science*, *47*, 1498–1514. https://doi.org/10.1287/mnsc.47.11.1498.10248

Bleichrodt, H., Schmidt, U., & Zank, H. (2009). Additive utility in prospect theory. *Management Science*, *55*(5), 863–873.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care*, *35*, 1095–1108. https://doi.org/10.1097/00005650-199711000-00002

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: Oxford University Press.

Hansen, K. S., & Østerdal, L. P. (2006). Models of quality-adjusted life years when health varies over time: Survey and analysis. *Journal of Economic Surveys*, *20*, 229–255. https://doi.org/10.1111/j.0950-0804.2006.00279.x

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., ... Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, *20*, 1727–1736. https://doi.org/10.1007/s11136-011-9903-x

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. https://doi.org/10.2307/1914185

Köbberling, V., & Wakker, P. P. (2005). An index of loss aversion. *Journal of Economic Theory*, *122*, 119–131. https://doi.org/10.1016/j.jet.2004.03.009

Martin, A. J., Glasziou, P., Simes, R., & Lumley, T. (2000). A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. *International Journal of Technology Assessment in Health Care*, *16*, 137–147. https://doi.org/10.1017/S0266462300161124

Miyamoto, J. M., & Eraker, S. A. (1988). A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General*, *117*, 3–20. https://doi.org/10.1037/0096-3445.117.1.3

Miyamoto, J. M., & Eraker, S. A. (1989). Parametric models of the utility of survival duration: Tests of axioms in a generic utility framework. *Organizational Behavior and Human Decision Processes*, *44*, 166–202. https://doi.org/10.1016/0749-5978(89)90024-1

Miyamoto, J. M., Wakker, P. P., Bleichrodt, H., & Peters, H. J. (1998). The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Science*, *44*, 839–849. https://doi.org/10.1287/mnsc.44.6.839

Noussair, C., Robin, S., & Ruffieux, B. (2004). Revealing consumers' willingness-to-pay: A comparison of the BDM mechanism and the Vickrey auction. *Journal of Economic Psychology*, *25*, 725–741. https://doi.org/10.1016/j.joep.2003.06.004

Oliver, A. (2003). The internal consistency of the standard gamble: Tests after adjusting for prospect theory. *Journal of Health Economics*, *22*, 659–674. https://doi.org/10.1016/S0167-6296(03)00023-7

Perpiñán, J. M. A., Martínez, F. I. S., Pérez, J. E. M. & Martínez, I. M. 2009. Debiasing EQ-5D tariffs. New estimations of the Spanish EQ-5D value set under nonexpected utility. Centro de Estudios Andaluces.

Pinto-Prades, J.-L., & Abellan-Perpiñan, J.-M. (2012). When normative and descriptive diverge: How to bridge the difference. *Social Choice and Welfare*, *38*, 569–584. https://doi.org/10.1007/s00355-012-0655-5

Pliskin, J. S., Shepard, D. S., & Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research*, *28*, 206–224. https://doi.org/10.1287/opre.28.1.206

Stiggelbout, A. M., Kiebert, G. M., Kievit, J., Leer, J.-W. H., Stoter, G., & De Haes, J. (1994). Utility assessment in cancer patients: Adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making*, *14*, 82–90. https://doi.org/10.1177/0272989X9401400110

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

Treadwell, J. R., & Lenert, L. A. (1999). Health values and prospect theory. *Medical Decision Making*, *19*, 344–352. https://doi.org/10.1177/0272989X9901900313

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.1007/BF00122574

van der Pol, M., & Roux, L. (2005). Time preference bias in time trade-off. *The European Journal of Health Economics*, *6*, 107–111. https://doi.org/10.1007/s10198-004-0265-y

van Nooten, F., & Brouwer, W. (2004). The influence of subjective expectations about length and quality of life on time trade-off answers. *Health Economics*, *13*, 819–823. https://doi.org/10.1002/hec.873

van Nooten, F., Koolman, X., & Brouwer, W. (2009). The influence of subjective life expectancy on health state valuations using a 10 year TTO. *Health Economics*, *18*, 549–558. https://doi.org/10.1002/hec.1385

van Osch, S. M., & Stiggelbout, A. M. (2008). The construction of standard gamble utilities. *Health Economics*, *17*, 31–40. https://doi.org/10.1002/hec.1235

van Osch, S. M., van den Hout, W. B., & Stiggelbout, A. M. (2006). Exploring the reference point in prospect theory: Gambles for length of life. *Medical Decision Making*, *26*, 338–346. https://doi.org/10.1177/0272989X06290484

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

Wakker, P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, *42*, 1131–1150. https://doi.org/10.1287/mnsc.42.8.1131

Wakker, P., & Stiggelbout, A. (1995). Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making*, *15*, 180–186. https://doi.org/10.1177/0272989X9501500212

Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*, 1329–1344. https://doi.org/10.1002/hec.1331

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity* Cambridge University Press. https://doi.org/10.1017/CBO9780511779329

SUPPORTING INFORMATION

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