QALYS without bias?

Non-parametric correction of time trade-off and standard gamble weights based on prospect theory

Stefan A. Lipman\textsuperscript{a}, Werner, B.F. Brouwer\textsuperscript{b} & Arthur E. Attema\textsuperscript{c}

\textsuperscript{a} (Corresponding author) Erasmus School of Health Policy & Management (ESHPM), Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, --31-10.4082507 (O), E: lipman@eshpm.eur.nl

\textsuperscript{b} ESHPM, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, E: brouwer@eshpm.eur.nl

\textsuperscript{c} ESHPM, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, E: attema@eshpm.eur.nl

Conflicts of interest: none

Funding source: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
QALYS WITHOUT BIAS?

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 modelled in prospect theory (PT): loss aversion and probability weighting. Furthermore, we correct for non-linear 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 non-parametric 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 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 demonstrate the importance of accounting for biases in health state valuations.

Keywords: health state valuation, time trade-off, standard gamble, utility of life duration, loss aversion, QALY model

Classification codes:
B41; D03; D81; I10

Acknowledgements: An earlier version of this paper was presented at the Lowlands Health Economics Study Group conference (Rotterdam, 2017), and the International Health Economics Association World Congress (Boston, 2017). We thank participants at both occasions for their comments. The authors would, furthermore, like to thank the following scholars for their valuable comments during the writing of this manuscript: Jan van Busschbach, Olivier l’Haridon and Han Bleichrodt. All remaining errors and bias are ours.
QALYS WITHOUT BIAS?

Introduction
In cost-utility analyses (CUA), the costs of medical technology are juxtaposed with the incremental health benefits, commonly expressed in Quality-Adjusted Life-Years (QALYs). The QALY model (Pliskin et al., 1980) entails multiplying prospective life years by weights, sometimes referred to as ‘utilities’, which represent health-related quality of life. These weights are normalized such that 0 represents the subjective weight or value of being dead and 1 a state without health problems. For this approach to be successful, it is imperative to validly determine the weights that are ascribed to certain health states; i.e., the QALY model, and therefore CUA, rely on empirically valid health state valuations. To this end, several methods are utilized to measure the values of quality of life, for the use in CUA, most notably Standard Gamble (SG) and Time Trade-Off (TTO). Empirical work on health state valuation, however, has demonstrated that health state weights are not independent of these two elicitation methods (e.g. Bleichrodt and Johannesson, 1997, Read et al., 1984, Torrance, 1976). Typically, SG weights, obtained through subjects’ decisions between staying in a less-than-perfect health and gambling for full health, are higher than TTO weights, which in turn are defined by the years of less-than-perfect health subjects are willing to trade off to obtain full health.

These discrepancies in elicited weights are often explained as the result of psychological biases, i.e. systematic errors which affect these supposedly equivalent methods dissimilarly (Bleichrodt, 2002). Indeed, an unresolved problem with the health state valuation using SG and TTO within the QALY model, is its sensitivity to such violations of expected utility (EU) theory, the classic theory of decision making under risk. For a review of these violations outside the health field, see the elaborate review by Starmer (2000). Such violations have also been documented within the field of health economics (Bleichrodt et al., 2007, Llewellyn-Thomas et al., 1982, Treadwell and Lenert, 1999). Additionally, the QALY model commonly assumes linear utility of life duration, which is equivalent to no discounting of future life years. Instead, many authors have found diminishing marginal utility of life years; i.e. life years that occur in the distant future tend to receive less weight than life years in the nearer future (e.g. Attema et al., 2012, Wakker and Deneffe, 1996, Bleichrodt and Pinto, 2005, Abellan-Perpinan et al., 2006). These findings question the use of the QALY model in its current form, and the conventional methods of valuation of health states, which also include a trade-off between quality of life and longevity. In order to better inform health-care decision makers, a methodological shift is required in health state valuation, towards the use of descriptive utility models such as prospect theory (PT).
Originally postulated as a descriptive alternative to EU, PT is characterized by four tenets (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). These are: (1) reference dependence – subjective value derived from a good is defined over differences from a reference point (RP), instead of over the overall consumption of that good; (2) loss aversion - the value function has an inflection point at the RP and is steeper for losses than for gains; (3) diminishing sensitivity - the value function is concave for gains and convex for losses, which indicates diminishing sensitivity to outcomes further from the RP; and (4) probability weighting - the decision maker overweights small probabilities, underweights large probabilities, and gives special attention to extreme probabilities (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992). Decades after its introduction, PT is well specified, and scholars have formally measured it for decisions with regard to income (Abdellaoui, 2002, Fehr-Duda et al., 2006, Tversky and Kahneman, 1992, Van De Kuilen and Wakker, 2011, Wu and Gonzalez, 1999) and for health outcomes (Bleichrodt and Pinto, 2000, Miyamoto and Eraker, 1989, Verhoef et al., 1994). Importantly, the biases modelled in PT affect the TTO and SG method differently, with loss aversion exerting an upward bias on both methods, but non-linear utility of life duration only affecting TTO while probability weighting only affects SG (Bleichrodt, 2002). Therefore, some studies have attempted to correct for these biases, in order to reduce discrepancies between health state valuation techniques (Perpiñán et al., 2009, Attema and Brouwer, 2009, Martin et al., 2000, Oliver, 2003, Wakker and Stigglerbout, 1995, Stigglerbout et al., 1994, van Osch et al., 2004).

Previous attempts at correcting TTO and/or SG weights for PT, however, suffered from several methodological problems, which are avoided in the present study. First, several studies did not correct for the full range of biases predicted by PT, but for example only loss aversion (e.g. Oliver, 2003) or non-linear utility of life duration (e.g. Attema and Brouwer, 2009, Stigglerbout et al., 1994). Second, earlier attempts at correction generally assumed a parametric form for utility of life duration and/or probability weighting, for example by fitting concave utility into the corset of a power or exponential function (Stigglerbout et al., 1994, van Osch et al., 2004, Martin et al., 2000, van der Pol and Roux, 2005). These practical yet simplifying procedures could have affected the estimates, thus allowing considerable bias to remain after correction (Abdellaoui et al., 2016, Abdellaoui et al., 2007). Third, earlier work may have underappreciated the role of the reference point (RP). Even though reference dependence appears to be the most central tenet of PT, earlier work in health economics on the location of the RP has produced heterogeneous results. More specifically, subjects have been suggested to utilize multiple different health outcomes as RP, such as: the lowest
possible outcome (Attema et al., 2012, Bleichrodt et al., 2001), the guaranteed outcome in TTO or SG (van Osch and Stiggelbout, 2008, van Osch et al., 2004), or the best outcome available (van Osch et al., 2006). Furthermore, the RP may be influenced by aspirations, expectations, norms, and social comparisons (Tversky and Kahneman, 1991).

We extend earlier work by utilizing recently developed non-parametric methodology (Abdellaoui et al., 2016), adapted to measure the utility of life duration instead of income. This method allows us to make PT completely visible in the health domain, and idiosyncratically correct weights for the biases subjects demonstrated in the TTO and SG methodology. Moreover, we attempt to append the heterogeneity surrounding RPs by providing subjects with an exogenous RP, a hypothetical vantage point in decision-making (following the successful procedures described in Attema et al., 2013, Robinson et al., 2001).

Our results imply that the present method can indeed be used to measure PT’s tenets in the health domain. We find considerable loss aversion, and probability weighting for both gains and losses. Furthermore, we find concave utility for life year gains and convex utility for losses. After correcting for these influences, the difference in weights between TTO and SG is no longer significant, demonstrating the importance of accounting for non-expected utility preferences in health state valuations.

The remainder of the paper is organized as follows: Section 2 covers notation and background, sections 3 covers the adapted methodology and the experimental procedure. In section 4 the results are presented, whilst section 5 features a discussion of these results. Finally, section 6 concludes.

2. Theoretical framework

Consider a decision maker facing choices with regard to his own health under uncertain conditions. We will consider lotteries about chronic health profiles, described as \((\beta, T)\), where \(\beta\) represents health status and \(T\) duration in years. For brevity, we will denote immediate death as \(D\) and if health status is equal to full health (FH) we will write \(\beta = FH\). Uncertainty is operationalized by presenting decision makers with various prospects which represent different life durations and quality of life. We let \((\beta_x, T_x)_p (\beta_y, T_y)\) denote the risky prospect that provides health profile \((\beta_x, T_x)\) with probability \(p\), and health profile \((\beta_y, T_y)\) with probability \(1 - p\). Also, we assume completeness, meaning that the decision maker has preferences over risky prospects, which are denoted using the conventional notation: \(>\), \(\geq\), and \(\sim\) to represent strict preference, weak preference, and indifference, respectively. Finally,
we assume monotonicity for both attributes. In other words, we assume that the decision maker prefers more health over less.

2.1. Linear QALY model, EU and health state valuation
Before turning to the background and notation relevant to our paper and the corrections we apply to TTO and SG, we first reiterate the conventional approach to obtaining TTO and SG weights, based on the linear QALY model and EU theory.

According to the linear QALY model the decision maker evaluates health profiles \((\beta, T)\) by:
\[
V(\beta, T) = U(\beta) \times T,
\]
with \(V(\beta, T)\) being a product of \(U(\beta)\), the utility of \(\beta\), and \(T\) life years. Within this framework, \(U(\beta)\) is obtained using either risky (SG) or riskless (TTO) prospects.

The SG method involves determining the probability \(p\) at which decision makers are indifferent between a sure outcome, described as \((\beta, T_S)\), and a risky prospect \((FH, T_S)_p(D)\).
In other words, QALY weights are determined by asking subjects to choose between a number of years \((T_S)\) in health state \(\beta\) for certain and a gamble with two outcomes, which are FH during the same time period \((T_S)\) and D. Typically, the probability \((p)\) of obtaining FH is varied until the respondent is indifferent between the two alternatives. Under EU, and assuming \(U(FH) = 1\) & \(U(D) = 0\), the indifference \((\beta, T_S)\sim(FH, T_S)_p(D)\) is evaluated by:
\[
1 \times U(\beta)T_S = p \times U(FH)T_S + (1 - p) \times 0,
\]
and thus:
\[
U(\beta) = p.
\]

The TTO method, on the other hand, asks for a time equivalent in perfect health which yields indifference between \((\beta, T_{T1})\) and \((FH, T_{T2})\), with \(T_{T1}>T_{T2}\). In other words, subjects are required to compare \(T_{T1}\) years in an impaired health state \(\beta\) to \(T_{T2}\) years in full health. The number of years in \(T_{T2}\) is varied until the respondent is indifferent between the two options.
Under the linear QALY framework, this indifference is evaluated by:
\[
U(\beta) = \frac{T_{T2}}{T_{T1}}.
\]

2.2. Generalized QALY model, PT and corrected health states
Now we turn to the notation relevant to the corrected TTO and SG weights. According to the generalized QALY model (Miyamoto and Eraker, 1989), a decision maker’s preferences for chronic health states, denoted by \((\beta, T)\), can be represented by the function:

\[
V(\beta, T) = U(\beta) \ast L(T),
\]

with \(L(T)\) denoting the utility of \(T\) life years. In other words, this model assumes the utility derived from quality of life and life duration to be separable, and to deviate from the linear QALY model by allowing for non-linear curvature of utility for life duration. Additionally, in this study we assume prospect theory (PT) with a sign-dependent utility function for life duration, with a separate evaluation of gains and losses in life duration.

We assume that through instruction, it is possible to set decision makers’ attribute-specific reference points (RPs) \((\beta_c, T_0)\) with \(L(T_0) = 0\). In order to elicit a continuous utility function for life duration, we elicit a standard sequence that runs through \(L(T_0) = 0\) while fixing quality of life at \(\beta_c\) throughout the whole task, where \(\beta_c\) represents a chronic health state. Thus, the RP for the non-parametric method is set at \((\beta_c, T_0)\). Outcomes that are strictly preferred to \((\beta_c, T_0)\) are defined as gains, whilst outcomes strictly less preferred to \((\beta_c, T_0)\) are defined as losses. Prospects involving both gains and losses are referred to as mixed prospects. Prospects are rank-ordered for duration, which indicates that for gain prospects the notation \((\beta_c, T_x)p(\beta_c, T_y)\) signifies that \(T_x \geq T_y\) and for losses it signifies that \(T_x \leq T_y\). For mixed prospects, on the other hand, the notation \((\beta_c, T_x)p(\beta_c, T_y)\) signifies that \(T_x\) is a gain and \(T_y\) is a loss.

Under PT with binary outcomes, the decision maker’s preferences over mixed prospects \(x_{pi}y\) are evaluated using a probability weighting function and a utility function:

\[
w^+(p)U(\beta_c)L^+(T_x) + w^-(1-p)U(\beta_c)L^-(T_y).
\]

while preferences over gain or loss prospects are evaluated by:

\[
w^i(p)U(\beta_c)L^i(T_x) + (1 - w^i(p))U(\beta_c)L^i(T_y), i = +, -.
\]

with \(i = +\) for gains and \(i = -\) for losses. \(L^i(T)\) is a standard ratio scale utility function, which is strictly increasing and real-valued with \(L^i(T_0) = 0\). Because of the interval properties of this scale, we can also fix the utility of one other outcome, which is dealt with by setting the utility of the lowest outcome equal to -1. We incorporate loss aversion by taking \(L^-(T) = \lambda L(T)\) for \(T < T_0\), where \(\lambda\) denotes a loss aversion index, with \(\lambda > 1\) \([\lambda = 1, \lambda < 1]\) indicating loss aversion [loss neutrality, gain seeking]. The probability weighting functions \(w^i, i = +, -\), assign a number to each probability, with \(w^i(0) = 0\) and \(w^i(1) = 1\). These probability weighting functions can be different for gains and losses and they are not
QALYS WITHOUT BIAS?

necessarily additive. Through the procedure described above, we only obtain \( w^i(\cdot) \) and \( \lambda \) alongside \( L(T) \). A more detailed description of how \( w^i(\cdot) \) and \( \lambda \) are determined can be found in Section 3.3 and 3.4.

2.2.1. SG weight correction
SG utility weights have been argued to be confounded by probability weighting and loss aversion, with both influences having an upward effect on weights (Bleichrodt, 2002). Before correction we need to set a life duration \( T_S \), that serves as vantage point. Earlier work, including using think-out-loud protocols (van Osch and Stiggelbout, 2008) suggested that the certain reduced health state \((\beta, T_S)\) usually functions as a RP in SG (Bleichrodt et al., 2001, Hershey and Schoemaker, 1985), which is also often framed as RP in our SG instructions. To utilize the elicited attribute-specific parameters for corrections, it is important to exogenously set the duration \( T_S = T_0 \), to ensure comparability in terms of the RP in longevity.

If we assume the reference point is \((\beta, T_S)\), the utility of this option is given by:

\[
U(\beta) \cdot L(T_0) = U(\beta) \cdot 0 = 0.
\]

The gamble \((FH, T_S)_p(D)\) then consists of the option to gain quality of life for \( T_S \) years with a probability \( p \) and a risk of immediate death with probability \( 1-p \), which is a loss in lifetime of \( T_S \) years in health state \( \beta \). The indifference \((\beta, T_S) \sim (FH, T_S)_p(D)\) can thus be evaluated by\(^1\):

\[
0 = w^+(p) \left( L(T_0) - L^-(0) \right) \ast (U(FH) - U(\beta)) + w^-(1-p) \lambda \left( L^-(0) - L(T_0) \right) \ast U(\beta). \tag{7}
\]

Substitution and solving for \( U(\beta) \) gives the corrected utility:

\[
U(\beta) = \frac{w^+(p)}{w^+(p) + \lambda w^-(1-p)}.
\tag{8}
\]

2.2.2. TTO weight correction
As described by Bleichrodt (2002), TTO utility weights are likely to be biased by decision makers’ non-linear utility function for life duration and by loss aversion. Typically, TTO exercises are framed with the impaired health state as RP \((\beta, T_{T1})\). We will not consider other

\(^1\) Note that we assume that our estimated probability weighting function also holds for quality of life improvements. Although it need not be the same, it is likely to be a good first approximation, since similar probability weighting functions, compatible with an inverse S-shape, have been reported for several different outcomes (Abdellaoui, 2000; Bleichrodt and Pinto, 2000; Attema et al., 2016).
RPs for the TTO, for multiple reasons. First, to our knowledge, no evidence exists to warrant the expectation that both outcomes in the TTO can serve as the RP. It seems unlikely that decision makers will utilize the full health outcome as their RP, as it changes in each iteration in choice-based procedures. Second, considering \( (T_{T2}, FH) \) would require estimation of loss aversion within the quality of life domain, which is beyond the scope of this paper. Consequently, we have to equate the change in utility of life duration as seen from the RP \((\beta, T_{T1})\) to the change in utility of health. The full health option of the TTO will be a gain in health status which is experienced for \( T_{T2} \) years, and a loss in lifetime \((T_{T1} - T_{T2})\) spent in the current health state. We, furthermore, adjust the TTO for loss aversion, by multiplying losses with \( \lambda \). Finally, to enhance comparability with SG, we set \( T_{T1} = T_0 \). Thus, the respondent is indifferent between the two options if the following equation holds:

\[
0 = (U(FH) - U(\beta))(L(T_{T2}) - L(0)) + \lambda^* (L(T_{T2}) - L(T_0))^* U(\beta). \tag{9}
\]

With \( U(FH) = 1 \) & \( L(T_0) = 0 \) & \( L(0) = -1 \), we obtain:

\[
(1 - U(\beta)) \times (L(T_{T2}) + 1) = -\lambda \times L(T_{T2}) \times U(\beta). \tag{10}
\]

Solving for \( U(\beta) \) gives:

\[
U(\beta) = \frac{L(T_{T2})+1}{(1-\lambda)L(T_{T2})+1}. \tag{11}
\]

3. Method

The present experiment consists of two parts, here referred to as utility elicitation and health state valuation. The first part (‘utility elicitation’) is based on the method proposed by Abdellaoui and colleagues (2016). The goal of our adaptation of this method is to non-parametrically correct conventional methodology used for QALY measurement ex post. As such, the second part of the experiment (‘health state valuation’) employs conventional methods of measuring QALY weights. We measure the utility of three less-than-perfect health states using both TTO and SG, in order to determine the utility of quality of life. We correct the values obtained through these two standard measures using the results of the first part of this experiment (see Eq. 8 and 11). This ex post process allows us to determine the corrected
utility of health states, after abating the influence of loss aversion, probability weighting, and non-linear utility of life duration

3.1. Experimental set-up
Subjects were 99 students of the Rotterdam School of Management (58 female), who participated in this experiment for course credits. The experimental session lasted for approximately 55 minutes and was run on computers in sessions of four subjects sitting adjacently in cubicles (they could not see each other). The experiment was run in Matlab, using the presentation software Psycho toolbox. An instructor was present at all times to answer any questions subjects might have with regard to the procedure. Subjects were encouraged to ask questions and explicitly asked to refrain from discussing with each other. Furthermore, they were told that there were no right or wrong answers and that they should go through the experiment at their own pace. After the computerized part, subjects filled out a paper-and-pencil questionnaire which measured several demographic and individual characteristics, such as age, gender and subjective life expectancy.

To reduce the influence of order effects, several counterbalancing procedures were conducted between participants. First, the order of the two parts of the experiment (utility elicitation vs. health state valuation) was counterbalanced. Within the utility elicitation part, the order in which participants faced gains or losses was randomized; half of the participants always completed gain sections first, whilst the other half completed loss sections first. Furthermore, within the health state valuation part, the order in which TTO and SG were presented was randomized, as was the order of the health states. Furthermore, a total of five practice blocks and four consistency checks were used (see Appendix A for details on where these were implemented).

3.2. Utility elicitation
The utility elicitation method used here was recently proposed by Abdellaoui and colleagues (2016). We adapted this methodology to measure PT under risk in the health domain, by focusing on the elicitation of the utility curve and the probability weighting function for life duration. More specifically, in our experiment utility elicitation consisted of four stages (an elaborate description of the method can be found in Appendix A). The first stage connects utility for gains to the utility for losses. The second and third stages employ 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 measures probability weighting, separately for
gains and losses. The present adapted methodology, thus, involves the determination of a standard sequence of outcomes (in terms of life duration), implying that the utility difference between successive elements in the sequence is constructed to be constant. As such, the method is distinctive from the original trade-off method, since the standard sequence features both gains and losses and runs through the reference point, allowing the estimation of loss aversion. This makes it possible to completely elucidate PT’s tenets in the health domain. In all cases, this method allows us to refrain from imposing parametric assumptions on utility and probability weighting.

Each of the four stages had slightly different instructions, providing the context for the trade-offs subjects were required to make. Across all stages, it was made clear to the subjects that they should imagine living until 70 years in perfect health, after which they would contract a disease and immediately die without any pain. This means that we fix the RP with $\beta_c = FH$ and $T_0 = 70 - \text{age of the subject}$ across all stages. Subjects had to choose between two drugs which could amend their situation. These drugs could have a degree of risk; the drugs could have a chance of being ineffective, or perhaps even have an adverse effect on health. Subjects were explained that the chance of success would be determined randomly, in other words it was specified that their personal characteristics had no influence on this chance. Additionally, subjects were explained that they would have to make a series of binary choices, after which they were required to specify an indifference value.

In each stage, the prospective outcomes were described before presenting subjects with a choice between drugs. Subjects were reminded that neither drug would influence their quality of life (as $\beta_c$ remained fixed). For stages featuring gains, subjects were explained that the drugs could have a beneficial effect on health, but could also be ineffective. Furthermore, as the standard sequence for gains involves mixed prospects (with the possibility of small losses), the drugs were explained to possibly have negative side effects, which could result in them living shorter than their RP $(FH, 70 - \text{age})$. For stages featuring losses, subjects were explained that in addition to the disease they were to contract at age 70, they suffered from another, unrelated, disease which had to be treated. This disease was to lead to losses in life duration (not affecting quality of life). However, some treatments provided an opportunity to decrease or remove this additional loss. This procedure allowed us to keep the RP at $(FH, 70 - \text{age})$.

In order to perform the elicitation described before, we had to specify a number of stimuli. The values chosen are depicted in Appendix B. The stimuli in the experiment featured deviations from the reference life expectancy, for example: a gain of 5 years amounts to
subjects living until 75, while a loss of -5 would signify living until 65. The indifference values for each prospect were elicited using a 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. The three binary choice questions corresponded to iterations of a bisection procedure. Additionally, the experiment did not accept slider values which violated stochastic dominance. This implies, for example, that if subjects faced prospect \( x_p 0 \), for instance, with \( x = 10 \), then any indifference value for a certainty equivalent, confirmed using the slider, should be smaller than 10. Similar procedures were in place for all indifference values elicited in the experiment (this holds for all indifferences described in Appendix B). If subjects provided such values they were presented with a warning message and requested to fill in a non-dominant value.

Furthermore, to avoid that the number of years traded in the TTO (maximum of 20) would exceed the scope of elicited utility curve, we had to ensure that subjects’ sequence continued to at least 20 years. If this was not the case, the corrections would require extrapolation beyond our measured curve, which we found undesirable. This variability in trade-off scale is unavoidable in the trade-off method, and was already discussed by Wakker & Denef (1996). Therefore, in this experiment, the standard sequence elicitation was only terminated when subjects produced an indifference value larger than 20 (or smaller than -20), or after 25 steps in the sequence to avoid extending the duration of the experiment beyond what was reasonable. Consequently, the number of elicitations in the standard sequence ranged from 4 to 25.

3.3. Health state valuation

For the valuation of less-than-perfect health states we used health state descriptions from the EQ-5D-5L (Herdman et al., 2011). The EQ-5D-5L distinguishes between five health domains, i.e., “mobility”, “self-care”, “usual activities”, “pain/discomfort”, and “anxiety/depression”. Within these domains, this taxonomy uses five health state levels from “no problems” to “extreme problems/unable to”. Different combinations of these five levels in each domain allow the EQ-5D-5L to describe over 3000 unique health states. In EQ-5D nomenclature, health states are represented by 5 digit codes like 22113. This example features as a label for a health state with: slight problems (i.e. level 2) with mobility and self-care, no problems with the usual activities and no pain/discomfort (i.e. level 1) and moderate anxiety/depression (i.e. level 3). A total of four health states were utilized in both the TTO
task and the SG task, one of which to practice these tasks. These health states were selected to reflect an array of mildly aversive health states, in order to avoid health states that could be considered worse than death. Such extreme health states are found to distort QALY measurement (Dolan, 1997), and are therefore beyond the scope of this paper. The following health states were used: 22222 (practice), 21211, 31221, and 32341, with the corresponding Dutch tariffs: 0.638 (practice), 0.876, 0.798, and 0.516 (van Hout et al., 2012, Versteegh et al., 2016). Hence, the results of three TTOs and three SGs are included in our analysis.

The TTO and SG were framed to subjects as a choice between treatments, instead of between drugs, in order to emphasize the transition to this distinctive part of the experiment. We applied a choice-based elicitation procedure with four choices. Previous research suggested that choice-based procedures produce more consistent measurements than direct matching (Attema and Brouwer, 2013, Bostic et al., 1990, Noussair et al., 2004). Subjects were asked to imagine living until their 50th birthday in perfect health, after which they contracted a disease which would affect their quality of life for their remaining life expectancy of 20 years. Subjects were to decide which of two treatments they preferred. In both cases the maximum expected life duration was set to 70 years - age, in order to retain attribute-specific RP for life duration across the different parts as described in the Background. In other words, in this second part subjects made decisions with regard to the quality of life for age 50 to 70 (with death following). By this procedure we tried to ensure that the durations in this part of the experiment were manageable for subjects, whilst maintaining $T_0 = 70 - age$ throughout the whole experiment.

For the SG task, subjects were explained that one of the treatments would allow them to live 20 years in a deteriorated health state ($\beta$), whilst the second treatment provided them with the opportunity to live the remaining 20 years in full health. This second treatment involved risk: if the treatment had an adverse effect, this would lead to immediate death at age 50. For TTOs, subjects were asked to imagine a choice between a treatment which would allow them to live their remaining life in a deteriorated health state, or living a shorter amount of time in a full health. Again, the time in reduced health state was kept constant at 20 years (from age 50 onwards), as earlier research on TTO and SG methodology has argued that the fixed outcome is often utilized as a RP (van Osch and Stiggelbout, 2008, van Osch et al., 2004). This assumption allows us to fix $T_0 = 70 - age$, as in the utility elicitation part of this experiment.

3.4. Analyses of loss aversion
Several definitions of loss aversion exist, with $\lambda$ being interpreted in various manners (see Köbberling and Wakker, 2005, Abdellaoui et al., 2007). Köbberling and Wakker (2005) provide a straightforward method to determine loss aversion, which is considerably easier than other definitions (i.e. Kahneman & Tversky’s (1979) definition). Under Köbberling and Wakker’s (2005) definition loss aversion ($\lambda$) is defined as the kink of utility at the reference point. That is, they define loss aversion as $U'(0)/U'_L(0)$, with $U'_L(0)$ representing the left derivative and $U'_R(0)$ the right derivative of $U$ at the reference point. To operationalize this definition, we computed each subject’s coefficient of loss aversion 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 $U(x_1^-)/x_1^+$ over $U(x_1^+)/x_1^-$, which is equal to $x_1^+/-x_1^-$ (see Abdellaoui et al., 2016 for a proof). 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 $x_1^+/-x_1^- < 1$.

3.5. Analyses of utility curvature
We used two methods to investigate utility curvature, a non-parametric and a parametric method (similar to the approaches used in Abdellaoui et al., 2016). For the first, non-parametric method, we calculated the area under the curve (AUC) of the utility function. The domain of $U$ was normalized to [0,1], for gains, by dividing all durations through the highest absolute duration answered by the subject in that domain ($x_{kG}^+$ or $x_{kG}^-$). If utility 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, while for losses the opposite direction holds (convex $>$ ½ , concave $<$ ½).

We also analyzed the utility function by estimating parametric functions, by nonlinear least squared. We employed the power family, as it is the most commonly employed parametric family. For this family, utility of life duration is defined by $x^\alpha$ with $\alpha > 0$. As is well known, for gains [losses] $\alpha > 1$ corresponds to convex [concave] utility, $\alpha = 1$ corresponds to linear utility, and $\alpha < 1$ corresponds to concave [convex] utility.

3.6. Probability weighting
We used certainty equivalences using varying probabilities to elicit probability weighting, similar to Attema and colleagues (2017). In particular, we used linear interpolation to obtain a probability weighting function for gains and losses, using $p=0.1, 0.3, 0.5, 0.7, 0.9$. Furthermore, we also employed Tversky and Kahneman’s one-parameter inverse S-shaped
probability weighting function \( w(p) = p^\gamma / (p^\gamma + (1 - p)^\gamma)^{1/\gamma} \), estimated by nonlinear least squares. The \( \gamma \)-parameter controls for the shape of the probability weighting function. If \( \gamma = 1 \) there is no probability transformation and \( w(p) = p \). 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 hold, corresponding to an S-shaped weighting function.

4. Results
Two subjects explicitly expressed unwillingness to trade off any life years, which caused the experiment to fail. These subjects were removed from further analyses. For the remaining 97 subjects we could determine a utility function for gains and losses. As can be seen in Appendix A, 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: p’s < 0.01). However, the correlations between these indifference values were significant both for gains and losses (Kendall’s τ’s > 0.63, p’s < 0.003). Furthermore, we found a difference for the consistency checks in the probability sequence for gains (paired t-test: p’s = 0.007), but not for losses (paired t-test: p’s = 0.62). The correlations between these values were also significant (Kendall’s τ’s > 0.51, p’s < 0.001).

Twenty-nine subjects violated monotonicity for health states, which indicates that they valued health states which were better or equal on each dimension lower than their dominated counterpart (e.g. 21211 vs. 31221). As we consider it is plausible that all subjects prefer more health to less, we reran the full analyses without these subjects and found no differences in the main results. As such, we report the results for the full sample below (n = 97).
4.1 Utility curvature for gains and losses in life years

In Figure 1 we plotted the median standard sequences of gains and losses together. The observed S-shaped pattern is consistent with the typical concave utility for gains and convex utility for losses under prospect theory (Kahneman and Tversky, 1979). Additionally, the utility function for gains appeared to be less steep than the utility function for losses, which indicates loss aversion. We replicated this pattern at the aggregate level, with median AUC for gains equal to 0.555, and for losses this non-parametric analysis produced a median AUC of 0.561, which were both significantly different from 0.5 (Wilcoxon signed ranks test: p < 0.001). After parametrically fitting a power utility function to the data, we found a median $\alpha$ of 0.787 for gains and 0.757 for losses (significantly smaller than 1, Wilcoxon signed ranks test: p < 0.001). Thus, both parametric and non-parametric results demonstrated concave utility for gains and convex utility for losses at the aggregate level.

Table 2 shows the classification of subjects’ utility curvature for gains and losses at the individual level, both parametrically and non-parametrically. We observe that the most common pattern was concave utility for gains and convex utility for losses, as was found in an earlier implementation of this method (Attema et al., 2017). Furthermore, this conclusion holds for both non-parametric (53%) and parametric (53%) results.
Table 2: Classification for utility curvature at the individual level

| Utility Curvature | Gains | Losses |
|-------------------|-------|--------|
|                   |       | Concave | Convex | Linear | Total |
| Non-parametric results |     |        |        |        |       |
| Concave           | 19    | 51     | 0      | 70     |
| Convex            | 7     | 17     | 1      | 25     |
| Linear            | 0     | 1      | 1      | 2      |

| Gains | Losses |
|-------|--------|
|       | Concave | Convex | Linear | Total |
| Parametric results |     |        |        |       |
| Concave           | 19    | 51     | 0      | 70     |
| Convex            | 6     | 18     | 1      | 25     |
| Linear            | 0     | 1      | 1      | 2      |

4.2 Loss aversion

Table 3 shows the results of our analyses of loss aversion. Utilizing Köbberling and Wakker’s (2005) definition, we found a median loss aversion index of $\lambda = 2$. Thus, we found considerable loss aversion at the aggregate level, with the median under being significantly higher than 1 (Wilcoxon test: p < 0.001). At the individual level, the majority of the classified subjects demonstrated loss aversion, with 72% loss averse subjects under Köbberling and Wakker’s (2005) definition.

Table 3: Aggregate statistics and individual classification for loss aversion

| Köbberling & Wakker’s (2005) definition | Median | IQR     |
|----------------------------------------|--------|---------|
| Loss averse                            | 70     | 15      |
| Loss neutral                           | 15     |         |
| Gain seeking                           | 12     |         |

4.3. Probability weighting

Figure 2 shows the median decision weights assigned to $p = 0.1, 0.3, 0.5, 0.7, 0.9$ using linear interpolation for gains and losses. As can been seen from the plots, we observed the typical inverse S-shaped probability weighting both for gains and losses, with more pronounced overweighting of small probabilities for losses. Using Tversky and Kahneman’s one-
parameter function, we found median $\gamma = 0.92$ for gains and median $\gamma = 0.84$ for losses (both significantly lower than 1, Wilcoxon test: $p$’s < 0.04). Both analyses demonstrated that the typical inverse S-shaped probability transformation was the most prevalent for our data, both for gains and losses. Moving to the individual level, for gains we found $\gamma < 1$ for 56 subjects (58%), $\gamma > 1$ for 41 subjects (42%). For losses we found more pronounced inverse S-shaped probability weighting, with 71 (73%), and 26 (27%), respectively.

**Figure 2: Probability weighting functions for gains and losses**

| Gains | Losses |
|-------|--------|
| p     | 0.1    | 0.3    | 0.5    | 0.7    | 0.9    | 0.1    | 0.3    | 0.5    | 0.7    | 0.9    |
| w(p)  | 0.19   | 0.34   | 0.45   | 0.62   | 0.76   | 0.31   | 0.46   | 0.56   | 0.71   | 0.81   |

4.4 Health state correction

Table 4 shows the raw and corrected weights for all health states using SG and TTO. Moreover, to test the sensitivity of our results to linear interpolation, we also corrected TTO and SG weights using a power utility estimate for life duration, and the Kahneman and Tversky probability weighting function (these are indicated by ‘Parametric Corrections’ in Table 4). As can be seen in Table 4, an initial difference in TTO and SG weights existed (paired t-tests, all $p$’s < 0.001), with SG weights being higher than TTO for all $\beta$. Our results show that the corrected weights were lower than the raw scores for TTO and SG (paired t-test: all $p$’s < 0.01). Out of the corrections reported here, the initially significant difference between the raw weights only fully disappeared for all $\beta$ after applying non-parametric corrections (paired t-test: all $p$’s > .09). The parametric corrections were unsuccessful in
alleviating the differences between weights, since large discrepancies between TTO and SG weights remained. When we tested these results for robustness by performing median (Wilcoxon) tests, we found similar results. The only exception was that the corrected difference for $\beta_3$ under non-parametric correction was significant.

Table 4: Overview of mean weights for health states $\beta_1$-$3$ for TTO and SG including differences between methodologies under multiple corrections.

| Health state | TTO weight | SG weight | Difference | TTO weight | SG weight | Difference |
|--------------|------------|-----------|------------|------------|-----------|------------|
| Raw weight   |            |           |            |            |           |            |
| $\beta_1$: 21211 | 0.665     | 0.75      | -0.085*** |            |           |            |
| $\beta_2$: 31221 | 0.605     | 0.706     | -0.101*** |            |           |            |
| $\beta_3$: 32341 | 0.39      | 0.518     | -0.128*** |            |           |            |
| Correction   | Non-parametric | Parametric |           |            |           |            |
| $\beta_1$: 21211 | 0.492     | 0.506      | -0.014n.s. | 0.496      | 0.598      | -0.102*** |
| $\beta_2$: 31221 | 0.442     | 0.456      | -0.014n.s. | 0.449      | 0.558      | -0.109*** |
| $\beta_3$: 32341 | 0.279     | 0.319      | -0.039n.s. | 0.295      | 0.387      | -0.092*** |

Note: *, **, and *** indicate that the differences were significant at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively, for paired t-tests.

As a final step, we isolated the effects of the biases we correct for one by one. For the sake of brevity, we only report these isolated corrections for the non-parametric corrections. In other words, we performed the successful corrections which are described in section 2.2.1 and 2.2.2 (Eq. 8 & Eq. 11), under four conditions. First, we corrected TTO for utility curvature only, with $\lambda = 1$. Second, TTO weights were corrected for loss aversion only, with linear utility. Third, we corrected SG for probability weighting only, with $\lambda = 1$. Finally, SG weights were corrected for loss aversion only, with $w(p) = p$. These corrections allow us to demonstrate the confounding influence of each bias in isolation. The mean weights after these isolated corrections can be found in Table 5 for TTO and Table 6 for SG. These results show that loss aversion had a stronger downward influence on TTO weights than utility curvature, and both probability weighting and loss aversion had a substantial negative influence on SG weights.
**Table 5: Isolated effects of corrections for utility curvature (UC) and loss aversion (LA) for TTO weights.**

|               | Raw score | Corrected score | UC only | LA only |
|---------------|-----------|-----------------|---------|---------|
|                | Implication: |                  |
| $\beta_1$: 21211 | $\lambda = 1$ | $U(T) = T$ |
|               | 0.665     | 0.492           | 0.611   | 0.537   |
| $\beta_2$: 31221 | 0.605     | 0.442           | 0.558   | 0.474   |
| $\beta_3$: 32341 | 0.39      | 0.279           | 0.364   | 0.288   |

**Table 6: Isolated effects of corrections for probability weighting (PW) and loss aversion (LA) for SG weights.**

|               | Raw score | Corrected score | PW only | LA only |
|---------------|-----------|-----------------|---------|---------|
|                | Implication: |                  |
| $\beta_1$: 21211 | $\lambda = 1$ | $w(p) = p$ |
|               | 0.75      | 0.506           | 0.643   | 0.63    |
| $\beta_2$: 31221 | 0.706     | 0.456           | 0.597   | 0.584   |
| $\beta_3$: 32341 | 0.518     | 0.319           | 0.459   | 0.387   |

**6. Discussion**

This paper provides evidence that Abdellaoui and colleagues’ (2016) method can be applied to the health domain, demonstrating a new way to correct the weights typically used in health state valuation, i.e. to reduce biases from TTO and SG. Through this experimental method, we estimate the full set of PT’s parameters in the health domain, in order to obtain more descriptively valid outcomes to be used within the QALY model. Our results are consistent with PT (Kahneman and Tversky, 1979): we observe concave utility for gains and convex utility for losses, inverse S-shaped probability weighting and considerable loss aversion.

Although earlier applications of PT in the health domain have occasionally found different results (e.g. Attema et al., 2013, Attema et al., 2016, Bleichrodt and Pinto, 2000, Bleichrodt and Pinto, 2005), we were able to remove the typical discrepancies between the TTO and SG method by corrections based on our data. The initial mean difference in weights observed for these relevant methodologies disappeared, after non-parametrically correcting for loss aversion, probability weighting and the convexity of life year losses.

The non-parametric method used in this paper was also applied in the health domain by Attema and colleagues (2017), in order to compare utility under risk to utility under ambiguity for health outcomes. The present work underlines this method’s relevance for health economics, by demonstrating that non-parametric corrections provide avenue to address one of the field’s long-standing problems, i.e. how to adequately value quality of life.
QALYS WITHOUT BIAS?

In general, the estimates of utility curvature for gains and loss aversion (when applicable) of earlier work are similar to ours (e.g. Attema et al., 2013, Attema et al., 2016, Bleichrodt and Pinto, 2000, Bleichrodt and Pinto, 2005), but differential results are found for the utility function for losses. These differences might be explained by methodological differences. For example, modifications of the trade-off method (as the one used here) might have a higher tendency to generate convex utility for losses than the semi-parametric method, which is a hypothesis that could be tested in future work. The application of Abdellaouï and colleagues’ (2016) method here is yet another demonstration that EU theory may systematically misrepresent actual preferences with regard to health outcomes (Bleichrodt et al., 2007, Llewellyn-Thomas et al., 1982, Treadwell and Lenert, 1999). This poses a problem to the conventional QALY framework, which is built on EU axioms (Pliskin et al., 1980). The method used in this paper may be a crucial step at appending this problem, with its potential to correct for all biases modeled in PT, without restricting these corrections to a specific parametric form.

Importantly, our results demonstrate that, through this newly developed methodology, it may be possible to abate the confounding influence of individual biases in health state valuations. We replicated the typical finding that uncorrected SG weights are higher than TTO weights. By means of corrections similar to those proposed by Bleichrodt and colleagues (2001), we attempted to remove the systematic bias in these weights, accountable to loss aversion, probability weighting and utility curvature. As a result, the weights assigned to both TTO and SG were markedly lower than their uncorrected counterparts, and no longer could differences be observed between the weights assigned to the three health states utilized here. Although successful attempts at correcting SG and TTO weights using parametric methodology are reported in earlier work (e.g. Stiggelbout et al., 1994, van Osch et al., 2004, Martin et al., 2000, van der Pol and Roux, 2005), our parametric corrections were not able to fully account for the discrepancies between these methods. When comparing the raw scores to the standard Dutch tariffs assigned to these health states, we observe large differences. After correction, this discrepancy gets even larger. For example, the Dutch tariff for health state $\beta_1$ (21211) is 0.876, while we elicited a raw TTO weight of 0.665, which even decreases to 0.492 after the non-parametric correction.

When we turn our attention to the effect of each bias in isolation, our results demonstrate that correcting for loss aversion and probability weighting (SG) has a more profound downward influence on weights than only correcting for utility curvature (TTO).
Contrasting these results to earlier attempts at correcting TTO and/or SG weights, we find similar results as van Osch and colleagues (2004), who find no effect of correcting TTO weights for utility curvature, and still observe discrepancies between TTO and SG after correcting the latter for probability weighting. Other authors did find effects of correcting TTO weights for non-linear utility, where such corrections increase TTO weights (e.g. Attema and Brouwer, 2009, Stiggelbout et al., 1994, Perpiñán et al., 2009, Martin et al., 2000). This strong contrast to the findings reported here is likely caused by the convexity we find for losses in life years, and the framing of the TTO and SG utilized in our design (which both featured losses in life years from the RP of living until 70 in a reduced health state). Future work could shed light on the degree to which this discrepancy may be caused by the non-parametric method or the framing used in our work.

Our results also have implications for earlier work on the role of RPs in health-related decision making. The heterogeneity with regard to the RP is well-known, consider for example that for health outcomes decision makers are argued to use one of multiple health outcomes as RP, e.g. the lowest outcome (Attema et al., 2012, Bleichrodt et al., 2001) or the guaranteed outcome (van Osch and Stiggelbout, 2008, van Osch et al., 2004). To append this heterogeneity, we instructed subjects explicitly to utilize an exogenous attribute-specific RP, which was fixed for life duration (70 years). 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 (reduced health state $\beta$) as RP. If the instruction was indeed successful, this finding would imply that subjects utilized the exogenous RP as status quo in their decisions, instead of their current (likely perfect) health. Although we do not directly test this assertion, our results, thus, imply that subjects can deliberately select different RPs as vantage points in health-related decision making, as was concluded in earlier work (Attema et al., 2013, Robinson et al., 2001).

Nevertheless, this study and its results should be considered in light of several limitations. First, some concerns may be raised with regard to the quality of our data. A considerable part of the subject pool violated dominance for the health states used here, assigning more weight to health states with more problems on at least one EQ-5D domain, all else being equal. Although they violated a seemingly vital part of rationality, excluding these subjects from the sample had no influence on our results. Considering it is unlikely that a third of our sample has masochistic preferences (i.e. prefer less quality of life over more), we expect that these errors in decision-making are to be attributed to lapses in attention, for example if subjects failed to notice differences in health states. A related limitation of our
study is the somewhat lower internal consistency observed in our consistency checks (for gains only), compared to earlier applications of this method (Attema et al., 2017, Abdellaoui et al., 2016). This discrepancy with earlier work may have been caused by our experimental environment, as subjects could freely specify any number on the sliders, instead of being limited to certain bounds, as in the original implementation of this method, giving subjects more opportunity to freely deviate from their earlier choices. We decided to incorporate this freedom to diminish the effects of error propagation, which refers to the process by which errors in early stages may profoundly affect later indifferences in chained methodology. In our adaption, subjects could rectify errors by adjusting the final indifference value on the slider to any non-dominant value in life years. As a final test, we performed an error simulation as described by Abdellaoui and colleagues (2005), which confirmed that errors did not have a propagating effect on the standard sequence we elicited for gains and losses. Furthermore, accounting for response error did not significantly affect our main conclusions, yielding similar weights for all health states. These TTO and SG weights did not differ significantly in all simulations for β1 and β2, whilst replicating our results in the majority of simulations for β3 (over 70%). These simulations demonstrate that this correction method is quite robust to error propagation.

Second, concerns may be raised about the role of the RP in this paper. We decided to induce an exogenous RP and, thus, cannot be certain that this RP was indeed constantly kept in mind by our subjects. Although the significant amount of loss aversion suggests that the induced RP impacted subjects’ decisions, we believe it is unlikely that subjects could perform this laborious process perfectly during the full experiment. Earlier work on health-related preferences has indeed demonstrated that individuals’ preferences are distorted by their own current health and life expectancy (Nooten and Brouwer, 2004, Van Nooten et al., 2009). This confounding effect may be caused by drops in subjects’ motivation or capability to engage in deliberate RP selection for each trade-off decision (i.e. depletion effects, see: Baumeister et al., 1998). However, in our work, we found no systematic association between subjects’ self-reported life expectancy and their respected estimates for loss aversion, utility curvature and probability weighting, nor were such associations observed for raw and corrected health state weights\(^2\). Nevertheless, the effect of subjective RPs remains an interesting follow-up question for further research. A related limitation concerns the assumption of attribute-specific RPs and how these were operationalized in this paper. Our operationalization hinges strongly on

\(^2\) All correlation coefficients demonstrated that these associations lacked statistical significance (all Kendall’s τ’s < 1.52, all p’s > 0.13).
the fixing of $T_0$ throughout the multiple parts of the experiment. This constant RP for life duration allowed us to correct for utility curvature, loss aversion and probability weighting, estimated within the same domain. We achieved this by assuming that subjects utilize the fixed outcome in both TTO and SG as their RP, while earlier work demonstrated that for SG subjects may also use full health as their RP (van Osch and 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 utilize full health as RP, both for TTO and SG.

Finally and perhaps most importantly, the results of the present research are to be interpreted with care, as the primary goal of the present research was merely to test the newly developed method’s applicability to correcting TTO and SG weights. An important etymological point to make is that we do not mean to imply that the corrected weights here are in fact ‘correct’. Although the typical difference between TTO and SG was no longer visible, the sizeable difference between the corrected weights and the Dutch tariffs (see van Hout et al., 2012, Versteegh et al., 2016) suggest that our approach needs further refinement. This marked difference could be the result of the nature of our sample, which consisted only of students, participating for course credit. Future research could increase sample size, and include a more representative cross-section of the general public, especially if this approach is to be used to obtain QALY tariffs for allocation decisions.

7. Conclusion
With the increasing importance of economic evaluations in health-related decision making, the question of how to determine the ‘true’ value of specific health states has become a crucial one. Indeed, the QALY model, and CUA equivalently, are reliant on empirically valid health state valuations. Conventional methodology, such as TTO and SG, have typically assigned different weights to the same health state, which has been explained as the result of psychological biases affecting each method differently (Bleichrodt, 2002). By means of Abdellaoui and colleagues’ (2016) non-parametric method we demonstrated that it may be possible to significantly reduce the effect of these biases in health state valuation. After correcting for loss aversion, probability weighting and utility curvature, TTO and SG weights for three health states are no longer discernable. Although it is tempting to imply that through the ex post corrections reported here, it may be possible to discern the ‘true’ value of specific health states, the markedly lower weights we found compared to the Dutch tariffs suggest that the method applied here should be tested more extensively within the health domain and
refined accordingly. This process may prove to be a crucial step in obtaining descriptively valid weights, i.e. to obtain QALY’s without bias. Although we advise against directly translating corrected weights to policy at this stage, our findings could be relevant for specialists utilizing CUA. Corrections based on PT, as utilized here, may significantly impact cost-effectiveness ratios, depending on the treatments being assessed and the health states pertaining to these medical technologies.

One might argue that through our corrections, psychological dimensions which are important to allocation decisions are artificially being removed. However, we believe that if decision makers want to account for these biases (if for example they believe that loss aversion is important in the evaluation of preventive measures) this would not be incompatible with the corrective approach applied here. We suggest that in such situations decisions makers would be better off not relying on weights with method-dependent biases, but should instead use the corrections suggested in this paper. If the context of the decision making problem would make them relevant, we advocate to separately weight back in psychological dimensions.
References

ABDELLAOUI, M. 2002. A Genuine Rank-Dependent Generalization of the Von Neumann-Morgenstern Expected Utility Theorem. *Econometrica*, 70, 717-736.

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.

ABDELLAOUI, M., BLEICHRODT, H. & PARASCHIV, C. 2007. Loss aversion under prospect theory: A parameter-free measurement. *Management Science*, 53, 1659-1674.

ABDELLAOUI, M., VOSSMANN, F. & WEBER, M. 2005. Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses Under Uncertainty. *Management Science*, 51, 1384-1399.

ABELLAN-PERPINAN, J. M., PINTO-PRADES, J. L., MENDEZ-MARTINEZ, I. & BADIA-LLACH, X. 2006. Towards a better QALY model. *Health Econ*, 15, 665-76.

ATTEMA, A. E., BLEICHRODT, H. & L’HARIDON, O. 2017. Measuring ambiguity preferences for health.

ATTEMA, A. E., BLEICHRODT, H. & WAKKER, P. P. 2012. A direct method for measuring discounting and QALYs more easily and reliably. *Med Decis Making*, 32, 583-93.

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.

ATTEMA, A. E. & BROUWER, W. B. 2013. In search of a preferred preference elicitation method: A test of the internal consistency of choice and matching tasks. *Journal of Economic Psychology*, 39, 126-140.

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.

ATTEMA, A. E., BROUWER, W. B., L’HARIDON, O. & PINTO, J. L. 2016. An elicitation of utility for quality of life under prospect theory. *Journal of health economics*, 48, 121-134.

BAUMEISTER, R. F., BRATSLAVSKY, E., MURAVEN, M. & TICE, D. M. 1998. Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology*, 74, 1252.

BLEICHRODT, H. 2002. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ*, 11, 447-56.

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.

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.

BLEICHRODT, H. & PINTO, J. L. 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management science*, 46, 1485-1496.

BLEICHRODT, H. & PINTO, J. L. 2005. The Validity of Qalys Under Non-expected Utility. *The Economic Journal*, 115, 533-550.

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.

BOSTIC, R., HERRNSTEIN, R. J. & LUCE, R. D. 1990. The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior & Organization*, 13, 193-212.
DOLAN, P. 1997. Modeling valuations for EuroQol health states. *Medical care*, 35, 1095-1108.

FEHR-DUDA, H., DE GENNARO, M. & SCHUBERT, R. 2006. Gender, Financial Risk, and Probability Weights. *Theory and Decision*, 60, 283-313.

HERDMAN, M., GUDEX, C., LLOYD, A., JANSSEN, M., KIND, P., PARKIN, D., BONSEL, G. & 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.

HERSHEY, J. C. & SCHOEMAKER, P. J. 1985. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science*, 31, 1213-1231.

KAHNEMAN, D. & TVERSKY, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263-291.

KÖBBERLING, V. & WAKKER, P. P. 2005. An index of loss aversion. *Journal of Economic Theory*, 122, 119-131.

LLEWELLYN-THOMAS, H., SUTHERLAND, H. J., TIBSHIRANI, R., CIAMPI, A., TILL, J. & BOYD, N. 1982. The measurement of patients' values in medicine. *Medical Decision Making*, 2, 449-462.

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.

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.

NOOTEN, F. V. & BROUWER, W. 2004. The influence of subjective expectations about length and quality of life on time trade-off answers. *Health economics*, 13, 819-823.

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.

OLIVER, A. 2003. The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Journal of health economics*, 22, 659-674.

PERPINÁN, 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.

PLISKIN, J. S., SHEPARD, D. S. & WEINSTEIN, M. C. 1980. Utility functions for life years and health status. *Operations research*, 28, 206-224.

READ, J. L., QUINN, R. J., BERWICK, D. M., FINEBERG, H. V. & WEINSTEIN, M. C. 1984. Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making*, 4, 315-329.

ROBINSON, A., LOOMES, G. & JONES-LEE, M. 2001. Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making*, 21, 17-27.

STARMER, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, 38, 332-382.

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.

TORRANCE, G. W. 1976. Toward a utility theory foundation for health status index models. *Health services research*, 11, 349.
QALYS WITHOUT BIAS?

TREADWELL, J. R. & LENERT, L. A. 1999. Health values and prospect theory. Medical Decision Making, 19, 344-352.

TVERSKY, A. & KAHNEMAN, D. 1991. Loss aversion in riskless choice: A reference-dependent model. The quarterly journal of economics, 106, 1039-1061.

TVERSKY, A. & KAHNEMAN, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5, 297-323.

VAN DE KUILEN, G. & WAKKER, P. P. 2011. The midweight method to measure attitudes toward risk and ambiguity. Management Science, 57, 582-598.

VAN DER POL, M. & ROUX, L. 2005. Time preference bias in time trade-off. The European Journal of Health Economics, 6, 107-111.

VAN HOUT, B., JANSSEN, M., FENG, Y.-S., KOHLMANN, T., BUSSCHBACH, J., GOLICKI, D., LLOYD, A., SCALONE, L., KIND, P. & PICKARD, A. S. 2012. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health, 15, 708-715.

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.

VAN OSCH, S. M. & STIGGELBOUT, A. M. 2008. The construction of standard gamble utilities. Health Econ, 17, 31-40.

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. Med Decis Making, 26, 338-46.

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. Med Decis Making, 24, 511-7.

VERHOEF, L. C., DE HAAN, A. F. & VAN DAAL, W. A. 1994. Risk attitude in gambles with years of life: empirical support for prospect theory. Medical Decision Making, 14, 194-200.

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.

WAKKER, P. & DENEFFE, D. 1996. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. Management science, 42, 1131-1150.

WAKKER, P. & STIGGELBOUT, A. 1995. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. Medical Decision Making, 15, 180-186.

WU, G. & GONZALEZ, R. 1999. Nonlinear Decision Weights in Choice Under Uncertainty. Management Science, 45, 74-85.
Appendix A: Overview of experimental parts and stages

| Part 1: Utility elicitation | Part 2: Health state valuation |
|-----------------------------|-------------------------------|
| Practice block 1: Certainty equivalents | Practice TTO (βp) |
| Stage 1: coupling gains and losses | TTO (β1, β2, β3) |
| Practice block 2: Trade-off method | Practice SG (βp) |
| Stage 2: Standard Sequence Gains | SG (β1, β2, β3) |
| Consistency block: Re-elicitation of $x_2^+$ | |
| Stage 3: Standard Sequence Losses | |
| Consistency block: Re-elicitation of $x_2^−$ | |
| Stage 4a: Probability weighting gains | |
| Consistency block: Re-elicitation of $x_{p1}^+$ | |
| Stage 4b: Probability weighting losses | |
| Consistency block: Re-elicitation of $x_{p1}^−$ | |
Appendix B: Elaborate description of measurement method

Abdellaoui and colleagues (2016) describe the following three-stage methodological procedure (see Table B1 for stimuli used in this adaptation of their method).

Consider a decision maker facing a choice with regard to life duration under uncertain conditions. Uncertainty is operationalized by presenting decision makers with various prospects involving real numbers, which represent different expected life durations. We let $x_p y$ denote the prospect that provides $x$ with probability $p$, and $y$ with probability $1 - p$. As such, we will refer to $x_p y$ as a risky prospect. We will assume monotonicity for life years, in other words, that the decision maker prefers more life years over less. Also, we assume the decision maker has preferences over risky prospects, which are denoted using the conventional notation: $\succ$, $\succeq$, and $\sim$ to denote strict preference, weak preference, and indifference, respectively.

Decision makers determine their preferences relative to an exogenous reference point $x_0$. Outcomes that are strictly preferred to $x_0$ are defined as gains, whilst outcomes strictly less preferred to $x_0$ are defined as losses. Therefore, prospects involving both gains and losses are referred to as mixed prospects. A gain prospect involves no losses (i.e. both $x$ and $y$ are at least weakly preferred to $x_0$) and, similarly, a loss prospect involve only losses. For gain prospects the notation $x_p y$ indicates that $x \geq y$ and for losses it signifies that $x \leq y$. In other words, the outcome that would deviate the most from the reference point $x_0$ is listed first. For mixed prospects, on the other hand, the notation $x_p y$ signifies that $x$ is the gain and $y$ is the loss.

Under prospect theory (PT) with binary outcomes, the decision maker’s preferences over mixed prospects $x_p y$ are evaluated using a probability weighting function and a utility function:

$$w^+(p)U(x) + w^-(1-p)U(y).$$  \hfill (A1a)

while preferences over gain or loss prospects are evaluated by:

$$w^i(p)U(x) + \left(1 - w^i(p)\right)U(y), i = +, -.$$  \hfill (A1b)

B.1. First stage: connection utility for gains to utility for losses

First, we select a probability $p$ that is kept constant throughout the first three stages and a gain $G$. Next, the indifference for loss $L$ is elicited, by providing subjects with the following choice $G_p L \sim x_0$. 

QALYS WITHOUT BIAS?

Equation (A1a) thus implies that:

$$w^+(p)U(G) + w^-(1-p)U(L) = U(x_0) = 0. \quad (A2)$$

Then, the certainty equivalents $x_1^+$ and $x_1^-$ are elicited, by the following indifferences:

$$x_1^+ \sim G_p x_0 \quad \text{and} \quad x_1^- \sim L_p x_0.$$  These indifferences consequently imply:

$$U(x_1^+) = w^+(p)U(G), \quad (A3)$$

and equivalently for losses:

$$U(x_1^-) = w^-(p)U(L). \quad (A4)$$

Combining Eqs. (A2) – (A4) gives:

$$U(x_1^+) = -U(x_1^-). \quad (A5)$$

Through this equation we obtain the first elements of the standard sequence ($x_1^+$ and $x_1^-$), which is elicited in subsequent stages.

**B.2. Second and third stage: elicitation of utility for gains and losses**

Next, the trade-off method by Wakker and Deneffe (1996) is employed to elicit a standard sequence. Let $\ell$ be a prespecified loss. First, subjects are presented with the prospects $x_1^+ p \mathcal{L}$ and $\ell_p x_0$, in order to elicit the loss $\mathcal{L}$ for which subjects are indifferent ($x_1^+$ is the gain from stage 1). The indifference $x_1^+ \sim \ell_p x_0$ gives:

$$w^+(p)U(x_1^+) + w^-(p)U(\mathcal{L}) = w^-(p)U(\ell). \quad (A6)$$

Through rearranging Eq. (7) we obtain,

$$U(x_1^+) - U(x_0) = \frac{w^-(p)}{w^+(p)}(U(\ell) - U(\mathcal{L})). \quad (A7)$$

Second, subjects are presented with the prospects $x_2^+ p \mathcal{L}$ and $x_1^+ \ell$, where the gain $x_2^+$ is varied such that they are indifferent $x_2^+ \sim x_1^+ \ell$. This indifference implies, after rearranging:

$$U(x_2^+) - U(x_1^+) = \frac{w^-(p)}{w^+(p)}(U(\ell) - U(\mathcal{L})). \quad (A8)$$

Combining Eqs. (A7) and (A8) gives:

$$U(x_2^+) - U(x_1^+) = U(x_1^+) - U(x_0). \quad (A9)$$

Finally, we elicit a series of indifferences, which together form the standard sequence for gains: $x_j^+ \sim x_{j-1}^+, \ell, j = 2, \ldots, k_G \{x_0, x_1^+, x_2^+, \ldots, x_{k_G}^+\}$. It follows straightforwardly that for all $j$, $U(x_j^+) - U(x_{j-1}^+) = U(x_1^+) - U(x_0)$.

The standard sequence for losses is constructed equivalently (part three). Subjects face similar prospects, where we first fix a gain $\mathcal{G}$ to elicit the gain $G$ such that subjects produce the
following indifference: \( G_p x_j^- \sim g_p x_0 \). Finally, we elicit the standard sequence \( \{ x_0, x_1^+, x_2^-, \ldots, x_{k_L}^- \} \) as described above using the following general form: \( G_p x_j^- \sim g_p x_{j-1}^-, j = 2, \ldots, k_L \).

### B.3. Fourth stage: probability weights

To measure the probability weighting functions \( w^+(p) \) and \( w^-(p) \), we asked for the certainty equivalents \( x_p^+ \) and \( x_p^- \) of the prospects \( x_{kG}^+ \) and \( x_{kL}^- \). The outcomes \( x_{kG}^+ \) and \( x_{kL}^- \) are the maximum (minimum) outcome elicited in the standard sequence. Therefore, it follows from the probability weighting function and the chosen scaling of utility that \( U(x_p^+) = w^+(p) \) and \( -U(x_p^-) = w^-(p) \). The values of \( U(x_p^+) \) and \( U(x_p^-) \) are interpolated from their respective standard sequences (elicited in stage 2 and 3). The probability \( p \) was varied \((0.1; 0.3; 0.5; 0.7; 0.9)\) to measure the probability weighting functions for a wide range of probabilities, both for gains and losses.

### Table B1: Four-stage procedure to measure utility of life duration

The third column shows the variables assessed in each stage, and column four shows the elicited indifferences. The fifth column shows the implication of these elicited indifferences. The final column shows the stimuli used in this experiment.

| Stimuli | Indifference | Implication |
|---------|--------------|-------------|
| \( L \) | \( G_p L \sim x_0 \) | \( U(x_j^+) = -U(x_j^-) \) |
| \( x_j^+ \) | \( x_j^+ \sim G_p x_0 \) | |
| \( x_j^- \) | \( x_j^- \sim L_p x_0 \) | |
| \( \ell \) | \( x_j^+ \ell \sim \ell_p x_0 \) | \( U(x_j^+) - U(x_j^-) = U(x_j^+) - U(0) \) |
| \( x_j^- \) | \( x_j^- \sim \ell x_j^- \) | \( -U(x_j^-) = U(x_j^-) - U(0) \) |
| \( g \) | \( g_p x_j^- \sim g_p x_{j-1}^- \) | |
| \( x_{p_i}^+ \) | \( x_{p_i}^+ \sim x_{p_i}^+ \) | \( U(x_{p_i}^+) / U(x_{kG}^+) = w^+(p) \) |
| \( x_{p_i}^- \) | \( x_{p_i}^- \sim x_{p_i}^- \) | \( U(x_{p_i}^-) / U(x_{kL}^-) = w^-(p) \) |