Testing the internal consistency of the lottery equivalence method: applications of expected utility theory and prospect theory

A master's thesis

Supervisor: Dr. A.E. Attema

Author: Paul Verhaak

Student number: 623352

Rotterdam, 28-07-2023

Word count: 8701

Table of contents

Abstract 3

List of abbreviations 4

1. Introduction 5

2. Theoretical framework 7

2.1. Current Research 7

2.2. Lottery equivalence method 8

2.3. Internal consistency lottery equivalence method 9

2.4. Prospect Theory 10

3. Research method 12

3.1. Respondents 12

3.2. Questionnaire 13

3.3. Sample exclusion criteria 14

3.4. Statistical analysis 15

4. Results 15

4.1. Sample characteristics 15

4.2. Main findings 16

5. Discussion and conclusion 18

Appendix 20

References 27

Abstract

The lottery equivalence method is seen as an option to measure the value of health states under risk. However, previous research has shown that the method is internally inconsistent when chaining to the failure outcome. The normal format of the lottery equivalence method uses death as worst health state in their prospects. Values generated via this method are called ‘basic reference values’. Another format of the method replaces death by another health state. Utility values generated through this format are called ‘chained to the failure values’. These values generally exceed ‘basic reference values’. Internal inconsistency is a problem, because it is unclear which format produces the correct utility value. This paper aims to improve previous research on several aspects in order to test whether internal inconsistency is present and what the possible cause of this internal inconsistency could be. The first improvement entails that a choice-based procedure will be used instead of a matching-based procedure in order to elicit utility values. Secondly, the probabilities of the gamble one of the gambles in the lottery equivalence method set at 0.5 for both health states present in that gamble instead of a probability of 0.9 for one health state of occurring and 0.1 for the other. Thirdly, health states are included in the survey instead of life years. Lastly, prospect theory is applied with the values of Tversky & Kahneman in order to possibly improve internal consistency of the method. In this study, under both expected utility as well as prospect theory with a fixed reference point for all participants and values of Tversky & Kahneman, the method has been found internally inconsistent, with ‘chained to the failure values’ exceeding ‘basic reference values’. These findings are in line with previous research on this subject. Although prospect theory as applied in this research has shown that the lottery equivalence method is internally inconsistent, the role of this theory in explaining people’s responses of the lottery equivalence method remains unknown, as application of this theory could explain people’s behaviour in relation to this method when it is sophisticated further. First, it is possible to measure loss aversion and probability weighting at the individual level instead of using the parameters of Tversky & Kahneman. Second, a reference point can be induced or measured at the individual level.

List of abbreviations

In this study, a set of abbreviations will be used repeatedly, hereby an overview.

Abbreviation	Meaning
EU	Expected utility
EUT	Expected utility theory
HSU	Health state utility
LE	Lottery equivalence method
PT	Prospect Theory
SG	Standard Gamble
PLE	Probability lottery equivalence method
RP	Reference point

1. Introduction

Healthcare resources are scarce, and in order to optimally distribute the available resources within the healthcare system, important trade-offs have to be made. Cost-effectiveness analysis is a rational manner on which to base these trade-offs on (Kim & Basu, 2021). One of the important inputs in those cost-effectiveness models, are health state utilities (HSU), which denote the value of living in a certain health state for one year in a cardinal manner (Brazier et al., 2019).

One method that is often used to measure the value of a health state is the Standard Gamble (SG). In the most common format of this method, a participant is asked to evaluate a choice between two life scenario's. In the first scenario, the individual will live the rest of his life with a health problem. In the second scenario, the risky alternative, the individual has p% chance to live the rest of his life in full health and 1-p% chance to die. The individual has to identify the value (probability) 'p' for which he is indifferent between both scenario's. With this probability, the utility of the health problem can be calculated (Salomon, 2014). The SG is often seen as the best approach in measuring health state utilities, because medical decision making often occurs under risk, which is something the SG accounts for (Brazier & Ratcliffe, 2017).

The behavioural theory through which the answers on the Standard Gamble are commonly analysed, is expected utility theory (EUT). The reason for using this method lies in the fact that EUT is the dominant normative decision theory under risk (Brazier & Ratcliffe, 2017). The expected utility (EU) of a scenario is calculated by summing the utilities of the different possible health states in a scenario, after they have been multiplied by their probability (ranging from 0-1) of occurring (Brazier & Ratcliffe, 2017). Indifference between two scenarios is reached when their EU is equal. Since the utility of the health problem is the only unknown factor in the SG (utility of *death* and *full health* are scaled at 0 and 1 respectively), its utility can be calculated. However, utilities elicited in the SG under EUT have been found to be descriptively inaccurate (Bleichrodt, 2002). This descriptive inaccuracy of EUT results in inconsistencies in utility measurements. The method is externally inconsistent under EUT, which means that other utility elicitation methods show different utilities for the same health state under EUT (Bleichrodt et al., 2007). Moreover, the method is internally inconsistent when chaining to the failure outcome. The normal format of standard gamble uses *death* as worst health state in their risky prospect. Utility values generated via this method are called 'basic reference values'. Another format of the method replaces *death* by another health state. Utility values generated through this format are called 'chained to the failure values'. These values generally exceed 'basic reference values'. Internal inconsistency is a problem, because it is unclear which format produces the correct utility value (Oliver, 2004). Incorrect utility estimates lead to biased resource allocation and subsequent worse societal health (Bleichrodt, 2002).

in order to eliminate the inconsistencies resulting from the inaccuracy of EUT in describing people's behaviour, scholars have applied Prospect Theory (PT) to analyse answers of the SG (Bleichrodt et al., 2001; Bleichrodt et al., 2007; Oliver, 2003). PT is seen as a better descriptive theory of behaviour under risk than EUT. First, the theory considers that people weight same sized losses more heavily than same sized gains. Besides this, it considers that people weight probabilities instead of evaluating them linearly (Kahneman & Tversky, 1979). Applications of PT alleviated external inconsistencies between methods but could not alleviate internal consistency of the SG when chaining to the failure outcome (Bleichrodt et al., 2001; Bleichrodt et al., 2007; Oliver, 2003). In other words, the SG method is internally inconsistent when chaining to the failure outcome under both PT and EUT. This implies that no accurate utility measurements can be generated with the SG.

A plausible reason for the internal inconsistency seen in the SG is the so called 'certainty effect' (Kahneman & Tversky, 1979; Oliver, 2004). This effect implies that people overvalue certain outcomes compared to uncertain outcomes, which creates an upward bias of SG utilities (Bleichrodt, 2002). Although this effect could in theory be eliminated by application of PT, which accounts for the fact that people overvalue the certain outcome by correcting for loss aversion, using the standard inputs for the PT formula found by Tversky & Kahneman (1992), has not been proven to be appropriate (Oliver, 2004).

Another method involving a component of risk that can be used to measure the value of health states is the lottery equivalence method (LE). The main difference of the LE compared to the SG entails the fact that a risky life scenario is substituted for the certain life scenario. This implies that the LE is a risk-risk method of eliciting health state utilities and the SG is a risk-riskless method (McCord & De Neufville, 1986). Since there is no possibility for a certainty effect in the LE, this method is seen as an alternative to the SG where internal consistency may hold, either under EUT or PT with the parameters of Tversky & Kahneman (1992).

Previous research (Oliver, 2005) on the internal consistency of the LE, possesses a couple of flaws and deficiencies that this study aims to solve. First, previous research calculated utility based of life duration. The main aim of health state valuation methods is to calculate the value of life for an individual as a consequence of life duration and his health status (i.e. the severity of a health state), so it may be useful to measure the internal consistency of the LE when health status is used. In other words, Secondly, the procedure that was used to elicit the utility in the previous study (Oliver, 2005) is a matching-based procedure. This procedure is seen as inferior to a choice-based procedure in measuring health state utilities (Attema & Brouwer, 2013; Bostic et al., 1990). Thirdly, in the lottery equivalence method, a participant is faced with a choice between two gambles, each with two health states. In one of those gambles the probability of occurrence of each health state is fixed beforehand by the researcher. Previous research (Oliver, 2005) fixed the probability of occurrence at 0.9 for the best health state in that gamble, and 0.1 for the worst health state in that gamble. The probability of occurrence for the best health state in this gamble is called 'q', and for the worst health state it is called '1-q'. Note that since there are only two health states in the gamble, 'q' + '1-q' must equal 1. However, there is a risk that the probability of occurrence for a health state of 0.9 influences participants' perception of the task. Namely, this gamble could be interpreted as a certain outcome for the health state with 0.9 probability of occurrence (Oliver, 2005). Lastly, the study only corrected for loss aversion and not for probability weighting, thereby forgoing the chance to improve internal consistency by applying prospect theory with the available values of Tversky & Kahneman (1992).

This research will address each these flaws separately. First, calculation of utilities will be based on health status instead of life duration. Secondly, a choice-based based procedure will be used instead of a matching procedure in this research in order to elicit utility values. Thirdly, 'q' is set equal to 0.5. Last of all, besides EUT, PT will be used to analyse the answers with the parameters found by Tversky & Kahneman (1992).

Previous research (Oliver, 2004; Oliver, 2005) testing the internal consistency of the SG and LE has shown internal consistency to be a much greater problem when chaining to the failure outcome than when chaining to the success outcome, therefore this research will focus just on internal consistency related to chaining to the failure outcome. Considering the improvements that must be made in relation to the previous research, our main research question will be the following:

Is the lottery equivalence method internally consistent when chaining to the failure outcome, when probability 'q' is equal to 0.5 and health states are expressed in terms of severity?

Our secondary research aim is to assess the extent of the presence of loss aversion and probability weighting that exists in people's responses to the LE questions. It is beyond the scope of this research to measure the degree of loss aversion and probability weighting for everyone separately and to test whether internal consistency of the LE improves when applying the individual level parameters. It is, however, possible, to apply PT with the predetermined values of Tversky & Kahneman, and to investigate whether the internal consistency of the method improves substantially compared to analysing answers under EUT. If internal consistency significantly improves under PT, this indicates that loss aversion and probability weighting affect the answers to LE questions. Our secondary research aim will therefore be the following:

Are utility estimates obtained by the lottery equivalence method biased due to loss aversion and probability weighting?

The paper will be structured as follows. In chapter 2, the current evidence surrounding the internal consistency of the LE will be discussed more deeply. Besides this, the structure of the LE will be outlined, and it will be explained how PT and EUT can be used to analyse answers of participants on the LE. In chapter 3, the form of the survey will be outlined. Moreover, the sample selection process, exclusion criteria, and the statistical methods used will be explained. In chapter 4, the results will be shown and explained, and the data quality will be discussed. Chapter 5 concludes.

2. Theoretical framework

2.1. Current Research

The SG, the gold standard for utility elicitation under risk, suffers from many problems when analysed through EUT, such as probability weighting, loss aversion, scale compatibility, that bias the results of this elicitation procedure (Bleichrodt, 2002). These biases explain a part of the internal inconsistency found in the SG (Oliver, 2003). Inconsistencies of the SG found under EUT, can partly be solved through application of PT, because this method accounts for probability weighting and loss aversion. For example, Bleichrodt et al. (2001) found that differences between elicitation methods, including 2 forms of the standard gamble, were removed when applying prospect theory which included the parameters proposed in the article written by Tversky & Kahneman (1992). Moreover, Oliver (2004) found that partial application of prospect theory, when only accounting for loss aversion and not for probability weighting, improved the internal consistency of the standard gamble substantially. Besides the fact that PT solves some problems related to the SG, the usage of the general structure of prospect theory, including probability weighting and the estimation of parameters by Tversky & Kahneman (1992), does not seem a reliable tool in eliminating the internal inconsistency of the SG completely (Oliver, 2003).

If no internal consistency is found in the LE under EUT, PT is applied with different reference points (RP). RP is a fundamental concept of PT that will be discussed in more detail later in this chapter. Shortly elaborating, outcomes better than the RP are evaluated as gains and outcomes worse than that reference point are evaluated as losses (Kahneman & Tversky, 1979). First, when application of PT to the answers of the LE substantially increases internal consistency, it is likely that the combination of probability weighting and loss aversion is present as a bias in the LE method. Besides this, it is unknown which reference point people assume in relation to the LE method, whereas it is clear in the SG (Bleichrodt et al., 2007). If the LE is internally consistent when a particular RP is assumed for all individuals when applying PT, this gives insight into which RP is used by people in response to answering LE questions. Lastly, if application of PT, with the original parameters proposed by Tversky & Kahneman (1992), reduces internal consistency of the PLE substantially, it gives support for the idea that PT in this format may be applicable to reduce biases created by the LE in general.

2.2. Lottery equivalence method

The main model of our research, on which the questionnaire and the results will be based, is the probability lottery equivalence method (PLE). The main structure of the PLE is shown in figure 1. The goal of the PLE is to elicit the health state utility value of X_2. X_1, X_2, X_3 are all different pregiven health states, with the preference ordering $X_1 \succcurlyeq X_2 \succcurlyeq X_3$, which entails that X_1 is preferred to X_2, which is preferred to X_3 (Bleichrodt, 2002; Oliver, 2005). Furthermore, the probability 'q' is fixed by the researchers and thereby also the probability '1-q', where 'q' can range from 0-1, as it is a probability. 'A' and 'B' are the two different treatments in the PLE. The question that is asked to the respondent is to give the probability 'p', ranging from 0-1, for which he/she would be indifferent between treatment A and B. In other words, the respondent should give the probability 'p' for which she/he is indifferent between gamble $(X_1, p; X_3, 1-p)$ and gamble $(X_2, q; X_3, 1-q)$ (Oliver, 2005). The gamble in which the probabilities are being kept fixed (i.e. the gamble with probabilities 'q' and '1-q') is called 'the stimulus' (Rodríguez-Míguez et al., 2019).

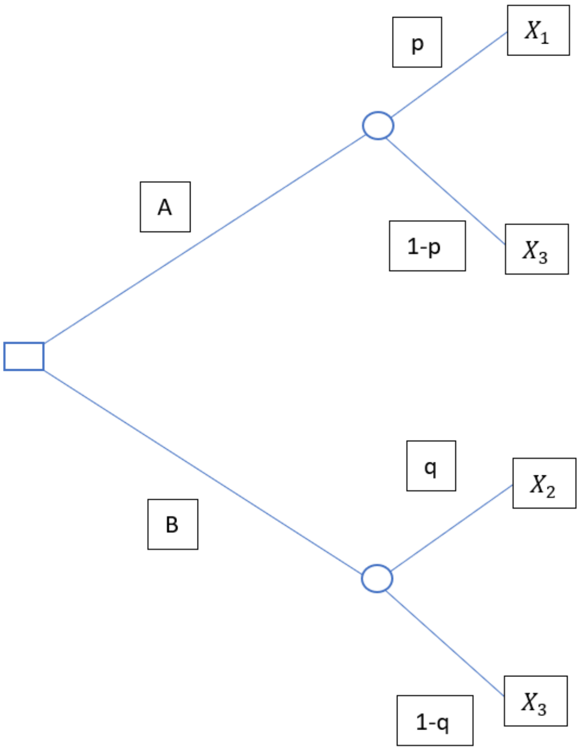


Figure 1. The main structure of the probability lottery equivalence method.

The main theory through which the outcomes of the PLE will be analysed is EUT, which is also the theory used mostly in the literature in analysing outcomes of PLE (Law et al., 1998; Oliver, 2005; Rodríguez-Míguez et al., 2019). In the context of the PLE, under EUT, expected utility of 'treatment A' should be equal to expected utility of 'treatment B', if one is indifferent between both treatments. When denoting u(.) as the value function of the PLE, the utility value of X_2 can be determined by the following formula when EUT is applied (Oliver, 2005) (see appendix 1 for derivation):

(1)

$$U(X_2) = \frac{pU(X_1) + (q-p)U(X_3)}{q}$$

It must be noted that the utility values of $u(X_1)$ and $u(X_3)$ are pregiven, so that the actual utility of X_2 can be calculated. The common way a utility value of a health state can be elicited using the PLE is through the basic probability lottery equivalence method. In this method, the state X_3 is set equal to *death* with a utility value of 0 and the state X_1 equal to full health with a utility value equal to 1. Utility values generated from this method, are called 'basic reference (lottery equivalence) values'. Another way the utility value of a health state can be elicited is through a PLE chained to the failure outcome. In this method, X_3 is substituted by a health state that is more severe than X_2 but less severe than the health state *death*. The utility value of this health state is 'chained in' from another basic probability lottery equivalence method (with *death* and full health present as X_3 and X_2 respectively), and with this chained in basic lottery equivalence reference value, a utility value of health state X_2 can be generated. Utility values generated through this method, are called 'chained to failure values'. In similar vein, the PLE can also be chained to the success outcome. In this method, X_1 is substituted by a health state worse than full health, but better than X_2. Again, the utility value of this health state is calculated in a basic probability lottery equivalence method, and therefore the utility value of X_2 can be calculated. Utility values generated by this method are called 'chained to success values' (Oliver, 2005).

2.3. Internal consistency lottery equivalence method

For the PLE to accurately measure cardinal utilities, the method ought to be internally consistent. In simple terms, this means that the 'basic reference values' of any health state elicited through the PLE do not systematically differ from either the 'chained to failure values' or the 'chained to success values' of that health state. Internal inconsistency occurs when this systematic difference does occur. The presence of internal consistency in a utility valuation method is crucial, as inconsistency casts doubt about which elicited utility value represents the preferences of the person filling in the questionnaire (Oliver, 2005).

In order to elaborate the point of internal inconsistency, the steps associated with obtaining a chained to failure value for a health state with the PLE under EUT will be outlined below, as well as the way to obtain the basic reference value for this health state with the PLE under EUT. Then, it will be specified which condition has to hold for the PLE to be internally consistent when, in this case, chaining to the failure outcome. It must be noted that the probabilities in the 'stimuli' (i.e., q and 1-q) are equal throughout the following questions. Suppose an individual is asked a probability p, such that he is indifferent between gamble $(X_1, p; X_3, 1-p)$ and gamble $(X_2, q; X_3, 1-q)$, where X_1 and X_3 are full health and death respectively and normalized to 1 and 0 respectively. X_2 is a predefined health state better than death and worse than full health. Application of EUT and therefore formula (1) tells us that the utility of health state X_2 will be equal to $\frac{p}{q}$. All In all, the basic reference value of health state X_2 is elicited under the PLE and analysed through EUT.

Let's consider another health state X_4, whose value can be specified by the preference ordering $X_3 \prec X_4 \prec X_2$. In similar vein, an individual is asked a probability, in this case 'r', for which he is indifferent between gamble $(X_1, r; X_3, 1-r)$ and gamble $(X_4, q; X_3, 1-q)$. Application of EUT and formula (1) tell us that the utility of X_4 is equal to $\frac{r}{q}$.

With the results of the previous basic reference PLE of health state X_4, the utility value of this health state is known, and therefore the chained to failure value of health state X_2 can be calculated. The person is asked to give a probability, in this case 'z', for which he is indifferent between gamble $(X_1, z; X_4, 1-z)$ and gamble $(X_2, q; X_4, 1-q)$. If the utility generated by this question (when chaining to the failure outcome) is different than in the basic reference PLE of health state X_2, then the PLE is internally inconsistent. For internal consistency, probability 'z' should be equal to:

(2)

$$z = \frac{p-r}{1-\frac{r}{q}}$$

(see appendix 2 for derivation).

2.4. Prospect Theory

It must be noted however, that EUT is not an accurate descriptive theory of decision making under risk (Kahneman & Tversky, 1979). In order to account for violations of people's behavior in relation to expected utility, the authors propose another theory, termed prospect theory (PT), through which to describe people's behavior in response to risky choices. Even more than 30 years after the publication of Kahneman & Tversky (1979), PT is still considered the best theory to describe behaviour under risk in an experimental setting (Barberis, 2013).

PT transforms EUT regarding some key aspects. The first difference between PT and EUT entails the idea that, in PT, people value changes in health or welfare rather than the absolute outcome, as is posed by EUT. In PT, it is hypothesized that people choose a reference point (RP), and outcomes better than that reference point are evaluated as gains and outcomes worse than that reference point are evaluated as losses (Kahneman & Tversky, 1979). Although PT is seen as a good theory to describe behaviour under risk, it is argued that it is very hard to determine what the reference point of an individual actually is, which makes the use of this theory complex (Barberis, 2013). Secondly, in PT, gains and losses are valued differently. More specifically, the disutility of experiencing a loss is greater than the utility that is experienced of a same sized gain, to a degree that can't be explained by risk aversion under EUT alone (Kahneman & Tversky, 1979; Oliver, 2003). In other words, people are 'loss averse'. Thirdly, it is hypothesized that people are more impacted by a change in a gain or a loss if its closer to the reference point rather than farther away from that point. In order words, people exhibit decreased marginal sensitivity to gains and losses (Kahneman & Tversky, 1979). In figure 2, the application of the previously mentioned aspects of PT is shown in a value function. The reference point is the intersection between the y-axis and x-axis. Gains and losses are concave and convex respectively, denoting decreased marginal sensitivity to gains and losses. The line denoting the value for losses is steeper than the line denoting the value for gains implying loss aversion.

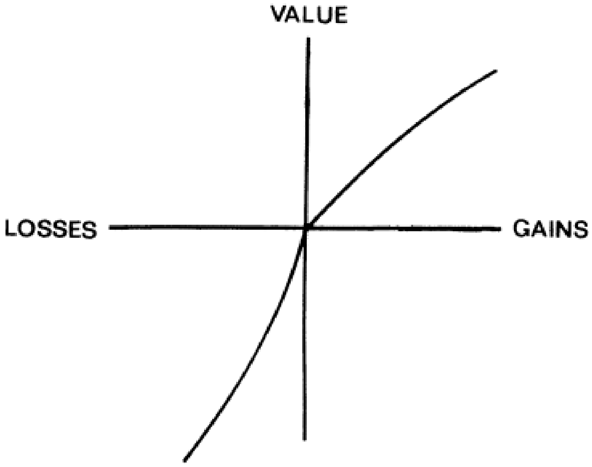


Figure 2. General value function in prospect theory (Kahneman & Tversky, 1979).

Lastly, PT, in contrast to EUT, accounts for the observation that people weight probabilities differently than their actual value (Kahneman & Tversky, 1979). It is often seen that people overweight small probabilities and underweight large probabilities (Bleichrodt, 2001). For example, if something has an objective chance of occurrence of 1%, overweighting of small probabilities implies that people place a higher weight on the occurrence of the outcome than 1%. They value it more than the value of the objective probability would suggest. The way this is operationalized in PT is that all probabilities in the formulas of PT are substituted by decision weights, which depend on the objective probability (Kahneman & Tversky, 1979). In figure 3, the common way probabilities are transformed into decision weights by people can be seen, where low probabilities are overweighted, and high probabilities are underweighted.

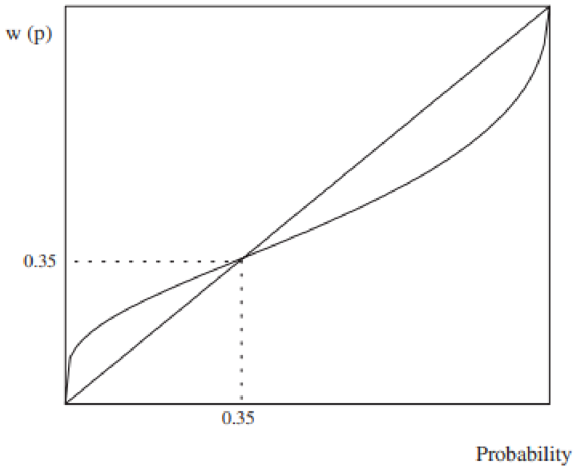


Figure 3. This figure shows the common probability weighting function, which shows how the probability weight (w(p)) is dependent on probability (Bleichrodt, 2002).

The general concept of PT can also be applied to the PLE. First, the probabilities that are present in the PLE must be weighted. This is done according to a probability weighting function proposed by Tversky & Kahneman (1992):

(3)

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}}$$

This function follows roughly the same shape as in figure 3. It must be noted that Tversky & Kahneman (1992) found different values for $'\gamma'$ for both losses and gains. $y^{+} = 0.61$ for gains, and $y^{-} = 0.69$ for losses, a study done by Bleichrodt & Pinto (2000) found roughly equal values for $y^{+}$and y^{-} in the field of health. The weighting function for gains is denoted as '$w^{+}(p)$' and for losses as '$w^{-}(p)$'. Secondly, in order to account for loss aversion in PT, a loss aversion parameter 'λ' is used which, denotes the relative value an individual places on a loss compared to an equally sized gain (Tversky & Kahneman, 1992). A value of $\lambda = 2.25$ was found by Tversky & Kahneman (1992).

In order to calculate utility values of the PLE under PT, a formula is needed to calculate the value of each gamble under PT. those formulas have been formulated by Bleichrodt et al. (2007) for gambles where only gains are present, only losses are present and where both gains and losses are present. After an individual has stated their indifference probability in the PLE, the value of health state X_2 can be found by equalizing the values of both gambles of the PLE. Solving for different reference points, the value of health state X_2 is characterized by the following calculations under prospect theory (see appendix 3 for an elaborate explanation):

(4)

For reference point worst health state (X_3) (i.e. only gains): $U(X_2) = \frac{w^{+}(p)+U(X_3)(w^{+}(q)-w^{+}(p))}{w^{+}(q)}$

(5)

For reference point full health (X_1) (i.e. only losses): $U(X_2) = \frac{1-w^{-}(1-p)+U(X_3)(w^{-}(1-p)-w^{-}(1-q))}{1-w^{-}(1-q)}$

(6)

For reference point intermediate health state (X_2) (i.e. both gains and losses): $U(X_2) =$

$$\frac{w^{+}(p)+U(X_3)(\lambda w^{-}(1-p)-\lambda w^{-}(1-q))}{w^{+}(p)+ \lambda(w^{-}(1-p)-w^{-}(1-q))}$$

The reason that three different formulas are used, is because it is unclear what the reference point of the respondents is. Reference points can, for example be, the current health state of an individual (Feeny & Eng, 2005) or the health state that is certain, when for example ´q´ in the PLE is equal to 1 (Oliver, 2003). The RP could also be the worst outcome (Bleichrodt et al., 2001), the most salient outcome (Oliver, 2005), or the best outcome (Van Osch et al., 2006). All in all, the RP is something specific to an individual. Although not all possible reference points can be explored in this study, the formulas applied cover a significant portion of them.

3. Research method

3.1. Respondents

Due to time and budget constraints, the total amount of respondents is relatively low. Moreover, convenience sampling was used to acquire participants (Stratton, 2021). Both factors limit the external validity of the study. Therefore, the research is exploratory in nature (Swedberg, 2020). The participants are recruited via two pathways. 37 participants in this study are drawn from the personal network of the researcher of this paper. An additional 39 participants were recruited via prolific.co, a website used by many scholars to acquire data. People in the personal network completed the study for free, whereas respondents recruited via prolific were gifted $2,50 upon completion of the survey. Prolific.co considers this reward to be very high given the fact that the survey is not too long, thus respondents were properly incentivized to take the survey seriously. By the researcher, a preselection criterium is applied to the respondents of Prolific which states that they must have completed a university degree. Although one may argue that such a criterium introduces selection bias, such a risk is outweighed by the fact that people who are, on average, smarter, should have a significantly higher chance to successfully fill out the survey because their understanding of the questions has a higher chance to be accurate.

3.2. Questionnaire

The questionnaire that will be filled in by the participants consists of three PLE questions: two basic reference PLE questions and one chained to the failure outcome PLE question. The questions are kept at a minimum, because a first pilot showed that a longer questionnaire was not feasible.

In this questionnaire, four health states need to be included to test the internal consistency of the method. An often-used method to describe health states s the EQ-5D-5L descriptive system (Versteegh et al., 2016). This system describes health states across the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, anxiety/depression. Those five dimensions all contain five levels of severity. A health state is constructed by choosing a level on each dimension, where '1' is used to denote the best outcome on a dimension and '5' is used to denote the worst outcome on a dimension. See appendix 4 for a more elaborate description of the levels of each dimension of the EQ-5D-5L descriptive system. Because the EQ-5D is a widely used tool to describe health states in health economics research, description of health states in this research will be based on the EQ-5D-5L (Versteegh et al., 2016).

In order to generate one chained to the failure outcome PLE value, two health states located in terms of preference, between death and full health, need to be generated. In order to generate valid conclusions about the internal consistency of the PLE method, the health states chosen need to satisfy four different conditions. First of all, health states that are valued closely to full health should be excluded. Namely, there is support for the fact that people do not want to increase their probability of dying in order to alleviate a health state that is very close to full health (Jones-Lee et al., 1995). Therefore, methods such as the PLE, a method in which the basic reference values are calculated whilst the probability of dying plays a pivotal role, may not be sensitive enough to calculate utility values for health states close to full health (Oliver, 2005). If those minor health states are used in the analysis regarding internal consistency, an outcome signifying internal inconstancy may be caused by insensitivity of the PLE regarding the assessment of the value of minor health states and not due to a fundamental problem with the method overall. Second, health states included in this research should not have a high chance to be valued worse than death by respondents, as the PLE in the current format does not allow health states to be valued worse than death. The third characteristic entails that there should be a clear preference relationship between the health states (i.e. which one is preferred to which). If it is not clear beforehand what the preference relationship between the health states is, it is impossible to assess how the chained to the failure values have to be generated. Due to the fact that the health states are defined as a set of levels across five different dimensions, lowering a level on one dimension (i.e. making that dimension better in terms of health) must imply that people prefer the changed health state, if it is assumed that people prefer more health, which is a reasonable assumption given classical microeconomic theory (Varian, 2010). Lastly, the health states should be relatively easy to interpret by the participants. PLE questions are hard to answer, therefore it is preferred to include health states in the questionnaire that only deviate from the health state full health on one or two dimensions according to the EQ-5D-5L rating system. Two health states that satisfy the conditions required are the health states ‘41111’ and ‘41113’ (Versteegh et al., 2016). The health states will be further elaborated in appendix 5.

Another important aspect of the questionnaire is to determine the probability ‘q’ of the stimulus gamble (see figure 1). Previous research on the internal consistency used a ‘q’ of 0.9 (Oliver, 2005). Using such a high probability is very risky, because the PLE has a high probability to be interpreted as a standard gamble (Oliver, 2005). In order to account for this problem, q is lowered to 0.5 in this research.

There are two general methods that can be used to elicit answers of the participants on the questions in the PLE. First of all, a matching task can be used, where the participant has to give a probability ‘p’ for which he would be indifferent between both gambles in the PLE (Attema & Brouwer, 2013). Another method is a choice task, in which an individual faces a series of choices where the value of ‘p’ varies and in each choice the participant must choose which gamble is preferred (Attema & Brouwer, 2013). After repeated choices for different values of ‘p’, an indifference value of the probability is elicited. Since matching-based procedures have been found to be less consistent than choice-based procedures, a choice-based procedure will be used (Attema & Brouwer, 2013; Bostic et al., 1990).

One such a choice-based elicitation procedure is the choice list methodology. This choice-based procedure has already been used to elicit preferences of health states (Attema et al., 2020; Arrieta et al., 2017). In this method, participants are faced with choices between two gambles. Based on a PLE structure of figure 1, the participant repeatedly must choose if he either prefers alternative A $'(X_1, p; X_3, 1-p)'$ or alternative B ‘$(X_2, q; X_3, 1-q)$’. Every variable is fixed throughout the questions except for the value of ‘p’, which starts at 0 (0%) for the first question and with increments of 0.1 (10%), ends at 0.5 (50%) for the last question. At a certain probability for p, the participant will shift his preference for alternative B to alternative A, e.g. when p goes from 20% to 30%. This implies that his indifference probability p of both gambles is somewhere between 20% and 30%. In order to figure out his exact indifference probability, a second choice list is presented to the participant. In the choice list for this participant, the conditions are similar, except for the fact that ‘p’ now starts at 0.2 and ends at 0.3, and the increments for ‘p’ are 0.02 (2%). In between two probabilities ‘p’, where the participant switches, lies the indifference probability (Attema et al., 2020). See appendix 6 for an example of such a procedure. The different elicitation questions that will be asked in the questionnaire will be randomized for each participant, to alleviate possible ordering effects, which entails that the order in which questions are answered may influence their result (Strack, 1992). The link to the full survey can be found in appendix 7.

3.3. Sample exclusion criteria

Respondents will be excluded from the dataset based on whether they pass four different tests. All those tests assume that people prefer more health to less health, which is a reasonable assumption based on microeconomic theory (Varian, 2010).

The first test consists of checking the final answers on the basic reference PLE questions of both health state A and B (see appendix 5). If the indifference probability ‘p’ of the basic reference PLE question for B is higher than for A, this implies that the individual prefers health state B to health state A, which is not logical and highly likely does not represent the individual’s true preferences but indicates that the individual did not understand the question.

The second test consists of checking whether respondents ‘switch back’ in the choice list. As ‘p’ of alternative A lowers throughout the choice list, alternative B becomes relatively more attractive. If the participant chooses to switch to alternative B at one point in the choice list, it is not logical for him to move back to alternative A as ‘p’ lowers, because alternative A gets relatively less attractive when this happens. Therefore, when a participant ‘switches back’ in the choice list question, he gets removed from the questionnaire. See appendix 8 of an example of a participant who switches back.

The third test consist of checking whether a participant prefers gamble $(X_2, p; X_3, 1-p)$ to gamble $(X_1, p; X_3, 1-p)$ when health state X_1 is better than X_2, which is also illogical based on the monotonicity assumption for health. Moreover, a fourth test is concluded to check whether a respondent preferred gamble $(X_1, 0; X_3, 1)$ to gamble $(X_2, 0.5; X_3, 0.5)$, when X_2 is better than X_3, thereby violating the monotonicity assumption of health.

3.4. Statistical analysis

Under EUT, one chained to failure value will be generated. This will happen for health state A. This chained to failure value will be compared to its matching basic reference value, which is also generated under EUT. Under PT, three chained to failure values will be generated, one for each reference point. The reference points are the worst health state in a gamble $'X_3'$, the intermediate health state in a gamble $'X_2'$ and full health $'X_1'$. Again, the chained to failure values will be calculated for health state A. For the three reference points, the basic reference values of health state A will also be calculated. Then, for each respective reference point, the basic reference values will be compared to the chained to failure values.

In order to test if there is a significant difference between the respective chained and basic reference values, a paired t-test will be used. The sampling distribution will very likely satisfy the normality assumption, even if the population distribution is nonnormal, due to the central limit theorem (Ott & Longnecker, 2015; Ross, 2020). If there is a significant difference between basic reference values and chained to the failure values, internal consistency is not present for the specification that is used to analyse the results. If the difference between the chained to failure values and the basic reference values are much smaller for a particular RP under PT than under EUT, there is a strong indication for the fact that loss aversion and probability weighting play a role in answering the PLE questions for participants.

4. Results

4.1. Sample characteristics

In total, 76 responses were recorded for the survey. Of those 76 responses, 26 did not met the criteria for usage in this research. 9 respondents were removed on the basis of the second exclusion criterion test, which implied that they switched back during one of their questions. An additional 5 respondents were removed on the basis of exclusion criterium test 3, and 5 more for test 4. Moreover, 7 respondents are removed from the survey because they violated test 1.

Table 1: Description of the samples.

	Unrestricted sample (n=76)	**Restricted sample (n=50)**
sex (male %)	44.00%	50.00 %
Average age, years	36.56	36.10
Recruited via prolific (%)	51.32%	52.00 %
Age distribution		
18-29	53.33%	52.00 %

30-44	14.67%	18.00 %
45-59	18.67%	20.00 %
60-71	13.33%	10.00 %
Level of education		
Completed secondary school	4.00%	6.00 %
Followed some university	12.00%	12.00 %
Completed bachelor's degree	60.00%	58.00 %
Completed a graduate or professional degree (MA, MS, MBA, JD, MD, DDS etc.)	24.00%	24.00 %

4.2. Main findings

In the research, three utility elicitation questions were answered by the participants. The basic reference PLE for health state A, the basic reference PLE for health state B, and the chained PLE for health state A through health state B. All the answers have been analysed through EUT and PT, for PT with reference points the intermediate health state, full health and worst health state present in their respective gamble. The formulas that are used to analyse the different answers are formula '(1)', '(3)', '(4)', '(5)' and '(6)' as described in the theoretical framework. Mean utilities and their respective standard deviation of the different PLE questions after application of the different theories can be found in table 2. Median utilities and their respective range are shown in table 3.

Table 2: Mean and standard deviation of utilities elicited

	Expected utility theory.	Prospect theory, reference point full health.	Prospect theory, reference point intermediate health state.	Prospect theory, reference point worst health state.
Basic reference A.	0.61 (0.24)	0.73 (0.22)	0.54 (0.22)	0.73 (0.22)
Basic reference B.	0.49 (0.29)	0.62 (0.30)	0.44 (0.25)	0.62 (0.29)
Chained A through B.	0.67 (0.28)	0.81 (0.22)	0.62 (0.24)	0.82 (0.21)

Mean utilities elicited and their respective standard errors for the three PLE questions calculated through various theories.

Table 3: Median and range of utilities elicited

	Expected utility theory.	Prospect theory, reference point full health.	Prospect theory, reference point intermediate health state.	Prospect theory, reference point worst health state.
Basic reference A.	0.66 (0.50-0.90)	0.79 (0.68-0.94)	0.57 (0.43-0.85)	0.79 (0.69-0.94)
Basic reference B.	0.56 (0.30-0.86)	0.73 (0.52-0.92)	0.48 (0.28-0.79)	0.73 (0.54-0.92)
Chained A through B.	0.77 (0.58-0.95)	0.91 (0.77-0.98)	0.69 (0.50-0.90)	0.91 (0.79-0.98)

Median utilities elicited and their respective interquartile range for the three PLE questions calculated through various theories.

The descriptive statistics show a couple of important characteristics of the utilities elicited in the sample. First, for all PLE questions, and for all theories, both the median and the mean chained utility value for health state A exceeds the basic reference value for that health state. In absence of further analysis, there seems to be a strong indication for internal inconsistency for the PLE for the health states studied and the theories that are used for analysis. Secondly, upon running the skewness test as proposed by D'Agostino and Belanger (1990), 10 of the 12 utility values elicited were found to be nonnormally distributed (p=0.05). Upon visually inspecting the utility distribution and by seeing that the median utility values exceed the mean utility values, the data has been found to be generally left-skewed. The last important observation, closely linked to the previous one, entails the fact that the answers on the PLE questions vary greatly. Although most people value living in health states A and B, some people would rather die.

A paired-t test is used in order to assess whether chained and basic reference values differ significantly from each other for the different theories (and reference points) that are used to analyse the answers. Pairwise differences of chained and basic reference values of health state A for different theories are shown in table 4 with their associated standard error shown in brackets and significance level shown by the number of stars. Chained PLE values have been found to significantly differ from direct chained PLE values for all theories in this research. Under expected utility theory, the difference is significant at a significance level of 5%. Under all three reference points of prospect theory, with the parameters used of Tversky & Kahneman (1992), the difference between the chained and basic reference value of health state A is significant at the 1% significance level. Prospect theory in this format did not provide any improvement for the internal consistency of the PLE over EUT in this research.

Table 4: paired t-test chained and basic reference values health state A (restricted sample, n=50).

Expected utility theory.	0.06 (0.028) **
Prospect theory, reference point full health.	0.08 (0.027) ***
Prospect theory, reference point intermediate health state.	0.08 (0.024) ***
Prospect theory, reference point worst health state.	0.09 (0.025) ***

Significance at 10% level = *, significance at 5% level = **, significance at 1% level = ***.

One could argue that the selection criteria for the sample are either too stringent or too lenient. Therefore, the sensitivity of the results in relation to the selection criteria will be tested. If people violate monotonicity concerns with regards to health within a PLE question (i.e., from two options they choose the worst one in terms of health), then this clearly demonstrates illogical behaviour that cannot realistically represent their preferences for health, as in response to a dilemma, they are unable to choose the correct option. Between PLE questions, monotonicity violations in terms of health could be argued to be reasonable, as people may change their minds during the time, they fill in the survey. Therefore, it could be argued that people who violate monotonicity concerns with regards to health between PLE questions should be included in the survey. Upon inclusion of those participants (n=57), the only meaningful difference that occurs compared to the restricted sample (n=50), entails the fact that internal inconsistency under EUT is significant at the 1% significance level instead of the 5% significance level (see appendix 9 for results). One could also provide a line of reasoning to further restrict the sample. Namely, some people have stated in the basic reference questions for health state A and B that they would rather die than live in such a health state. Because the PLE in the current format is not able to capture worse than death values, the value those participants put on those health states is recorded as equal to death in this survey. Their true valuation is unknown, and this makes their answers less useful. When those participants are excluded from the restricted sample, 45 participants remain. No meaningful changes in the results were found after application of these additional sample selection criteria when compared to the restricted sample (n=50; see appendix 10 for results). One could also argue that there is an additional criterion to exclude respondents. It could be argued that finishing the questionnaire within a timeframe that is too short indicates that it can't be filled out seriously. Based on the pilot that is run of the survey, the minimum time needed to fill in the survey seriously is estimated at 5 minutes. If participants are excluded from the restricted sample when they filled it out in less than 5 minutes, 39 participants remain. The results do not change. All in all, the results do not change meaningfully when accounting for various reasonable forms of restricting the sample.

5. Discussion and conclusion

This study confirms findings about the internal consistency of the LE under EUT of previous research when chaining to the failure outcome. After application of four distinct improvements to the research design compared to previous research, the results do not change (Oliver, 2005). Since EUT is considered a descriptively inaccurate theory to describe people's behavior under risk (Kahneman & Tversky, 1979), these findings do not come as a surprise. Under PT, the LE has also been found to be internally inconsistent in this research. Under both theories, chained values exceed basic reference values, which is also seen in previous research (Oliver, 2005). Whether PT is unable to correct problems regarding internal consistency remains unclear. Besides this, it remains unclear to what extent PT plays a role in explaining people's behavior in the LE. Therefore, it remains unclear to what extent probability weighting and loss aversion play a role in the LE method. Namely, the application of PT, a theory which accounts for loss aversion and probability weighting, in this paper is crude, and a more sophisticated approach could plausibly eliminate internal inconsistency of the LE.

Regarding two aspects, the application of PT in this paper can be classified as unsophisticated. First, probability weighting and loss aversion parameters found by Tversky & Kahneman (1992) and used in this paper are median values located in the money domain. Parameters focused on the money domain may differ significantly from parameters in the health domain (Attema et al., 2013). Moreover, using median values for all participants may give very unreliable results, as probability weighting and loss aversion parameters vary wildly between individuals (Abdellaoui, 2000; Brown et al., 2021; Gonzalez & Wu, 1999).

Secondly, the reference points of the respondents are not clear. In an optimal scenario, reference points of the individuals would be known, such that accurate assessment of the value of different health states based on prospect theory could be made. In this research, the same reference point is assumed for all respondents simultaneously for prospect theory analysis. This makes it unable to capture the real value of prospect theory, in the case in which the reference points differ significantly between participants. Reference points could be the current health state of the participant (Feeny & Eng, 2005), the worst outcome (Bleichrodt et al., 2001), the most salient outcome (Oliver, 2005), or the best outcome (Van Osch et al., 2006).

A limitation of this research is the manner of data collection in this research. Half of the data is collected via Prolific, a tool used by researchers to acquire a random sample of participants for their research. The other half of the data is collected through relatives who were willing to fill in the survey. In other words, convenience sampling was used for this part of the dataset. A risk of using this manner of sampling entails the fact that the sample is not representative of the population and that this may bias the results. This concern is significantly alleviated when we inspect the relationship between being recruited through prolific or through personal affiliation with the researcher and the degree of internal consistency present. When controlling for age (and age squared), educational attainment and gender, there is no relationship between the degree of internal consistency and being recruited through prolific compared to not being recruited through prolific (p value > 0.1) for the four theories. All in all, the lack of a representative sample provides a modest threat to the generalizability of the results.

A second limitation entails the concern that the sample size is too small. This threatens the validity of the results as decreased study power decreases the chance that a statistically significant effect reflects a true effect in a population (Button et al., 2013). Since internal inconsistency under PT has been found to be very significant in this study, those results are not affected too much by concerns of small sample size (Button et al., 2013). Internal inconsistency under EUT was not significant at the 1% level. Therefore, a small sample size poses a slight threat to the validity of the results for the internal consistency of the PLE under EUT.

A third limitation consists of the fact that there was no interviewer present during the time that the questions were answered by the participants. Difficulty in answering the questions by the participants is based on two facts. First, some relatives commented on the fact that the questions were quite difficult. Second, 34% of the sample was excluded of the analysis because their answers were illogical regarding the assumption of monotonicity of health. There is a slight chance that people who passed the tests regarding the selection criteria, still did not understand the questionnaire. However, the criteria were quite stringent, so that chance that they did not understand the questions is considered low. threat of misunderstanding of the questions on the validity of the results is considered moderate.

Although the paper contains some limitations, especially considering the absence of a researcher whilst the participants answer their questions, it is recommended, based on the results, in line with previous research (Oliver, 2005), to not use the PLE to elicit Cardinal Health state utilities under either EUT or PT as proposed in this paper. The standard gamble suffers from the same problem with regards to internal inconsistency as the lottery equivalence method when chaining to the failure outcome (Oliver, 2003) when applying prospect theory with the values of Tversky & Kahneman. However, external consistency is compromised for the lottery equivalence method and not for the standard gamble, when applying prospect theory with those values (Bleichrodt et al., 2007). Besides, questions related to the LE are more difficult to answer than the SG, because the LE is a risk-risk question, and the SG is a risk-riskless question. Due to these factors, the SG remains the gold standard for utility measurement under risk.

In line with previous research (Bleichrodt, 2002) on biases affecting health state utility measurement, the results suggest either one or two things. First, it could be possible that the application of PT used in this paper is not able to capture loss aversion, probability weighting and reference point in an accurate manner, because it is not measured at the individual level. Second, scale compatibility could cause internal consistency, when respondents overvalue certain probabilities in PLE questions. A combination of both factors as cause of internal consistency of the LE is also possible. All in all, when chaining to the failure outcome, the LE has been found to be internally consistent. Application of PT with parameters of Tversky & Kahneman (1992), has not been able to solve this.

Appendix

Appendix 1:

Suppose an individual is indifferent between gamble $(X_1, p; X_3, 1-p)$ and gamble $(X_2, q; X_3, 1-q)$. If an actor behaves according to EUT and if the utility of full health is set to 1 $'U(X_1) = 1'$, this implies that the EU of both gambles must be equal. This can be shown in the following equation:

$$p + U(X_3) * (1-p) = U(X_2) * q + U(X_3) * (1-q)$$

Solving for '$U(X_2)$' gives us the following derivation:

$$U(X_2) = \frac{pU(X_1) + (q-p)U(X_3)}{q}$$

Appendix 2:

Suppose an actor is indifferent between gamble $(X_1, z; X_4, 1-z)$ and gamble $(X_2, q; X_4, 1-q)$. From previous basic reference questions, it is already known that $'U(X_2)'$ is equal to $\frac{p}{q}$ and $'U(X_4)'$ is equal to $\frac{r}{q}$ under EUT. Then, the only way this new indifference can be consistent with the previous indifferences under EUT is in the following manner under EUT (utility of full health is set to 1):

$$1.\ z + U(X_4) * (1-z) = U(X_2) * q + U(X_4) * (1-q)$$

$$2.\ z + \frac{r}{q} * (1-z) = \frac{p}{q} * q + \frac{r}{q} * (1-q)$$

$$3.\ z + \frac{r}{q} - \frac{zr}{q} = p + \frac{r}{q} - r$$

$$4.\ z(1 + \frac{r}{q}) = p - r$$

$$5.\ z = \frac{p-r}{1-\frac{r}{q}}$$

Appendix 3:

For the PLE, the participant is asked to give their probability p for which he is indifferent between gamble $(X_1, p; X_3, 1-p)$ and gamble $(X_2, q; X_3, 1-q)$, where $X_1 \succcurlyeq X_2 \succcurlyeq X_3$ and X_1 is equal to full health. Health state full health and death are scaled to 1 and 0 respectively. Strictly speaking, a time element 'T' should be included in the structure of the gambles of the PLE, but for every derivation of the PLE, T is factored out, so we do not include it in the main text. For the derivations below, it will be included. Strictly speaking, the participant is asked his indifference probability for gamble $((X_1, T), p; (X_3, T), 1-p)$ and gamble $((X_2, T), q; (X_3, T), 1-q)$. Because the participant is asked his indifference probability, the values denoting the value of each gamble should be similar. The goal is therefore to find the utility value of X_2 for which both values of the gambles are similar for different reference points.

Gambles are evaluated differently depending on the reference point and if the outcomes in the gamble are gains or losses relative to that reference point (Bleichrodt et al., 2007). If a gamble involves both a gain and a loss relative to its reference point, it is evaluated in the following manner according to the formulas of PT used by Bleichrodt et al., (2007):

For the gamble $((X_1, T), p; (X_3, T), 1 - p)$, with the preference relationship $(X_1, T) \succ (X_{RP}, T) \succ (X_3, T)$, where (X_{RP}, T) is the reference point health state, the gamble is evaluated in the following manner under PT (Bleichrodt et al., 2007):

(7)

$$\begin{aligned} &\mathrm{PT}((X_1, T), p; (X_3, T), 1 - p) \\ &\quad = U(X_{RP}, T) + w^{+}(p)(U(X_1, T) - U(X_{RP}, T)) - \lambda w^{-}(1 - p)(U(X_{RP}, T) \\ &\quad - U(X_3, T)) \end{aligned}$$

For the gamble $((X_1, T), p; (X_3, T), 1 - p)$, with the preference relationship $(X_1, T) \succcurlyeq (X_3, T) \succcurlyeq (X_{RP}, T)$, where (X_{RP}, T) is the reference point health state, the gamble is evaluated in the following manner under PT (Bleichrodt et al., 2007):

(8)

$$\begin{aligned} &\mathrm{PT}((X_1, T), p; (X_3, T), 1 - p) \\ &\quad = U(X_{RP}, T) + w^{+}(p)(U(X_1, T) - U(X_{RP}, T)) + (1 - w^{+}(p))(U(X_3, T) \\ &\quad - U(X_{RP}, T)) \end{aligned}$$

For the gamble $((X_1, T), p; (X_3, T), 1 - p)$, with the preference relationship $(X_{RP}, T) \succcurlyeq (X_1, T) \succcurlyeq (X_3, T)$ where (X_{RP}, T) is the reference point health state, the gamble is evaluated in the following manner under PT (Bleichrodt et al., 2007):

(9)

$$\begin{aligned} &\mathrm{PT}((X_1, T), p; (X_3, T), 1 - p) \\ &\quad = U(X_{RP}, T) - \lambda w^{-}(1 - p)(U(X_{RP}, T) - U(X_3, T)) - \lambda(1 \\ &\quad - w^{-}(1 - p))(U(X_{RP}, T) - U(X_1, T)) \end{aligned}$$

For the PLE, the aim is to find the solution for the following equation: $((X_1, T), p; (X_3, T), 1 - p) = ((X_2, T), q; (X_3, T), 1 - q)$. In this research, three different reference points are chosen, (X_1, T), (X_2, T) and (X_3, T). The unknown that must be found is X_2. For each of those three reference points, the formula to arrive at the unknown value will be derived, given formula 7, 8 and 9. Besides this, it is assumed that the QALY model holds, which means the multiplicativity assumption holds which implies that the utility of a health state 'X' for duration 'T', formally denoted as $'U(X, T)'$, can be evaluated as $'U(X) * T^{r}{}'$. Empirical evidence has been found to support the multiplicativity assumption (Miyamoto & Eraker, 1988; Bleichrodt & Pinto, 2005).

For reference point $'(X_3, T)'$, the following equation can be solved by applying formula 7, scaling $'U(X_1)'$ to 1 and by assuming multiplicativity in the following manner $((X_2, T), q; (X_3, T), 1 - q) = ((X_1, T), p; (X_3, T), 1 - p)$:

1. $U(X_3) * T^{r} + w^{+}(q)(U(X_2) * T^{r} - U(X_3) * T^{r}) = U(X_3) * T^{r} + w^{+}(p)(U(X_1) * T^{r} - U(X_3) * T^{r})$
2. $w^{+}(q)(U(X_2) - w^{+}(q)U(X_3) = w^{+}(p)U(X_1) - w^{+}(p)U(X_3)$
3. $w^{+}(q)(U(X_2) = w^{+}(p) + U(X_3)(w^{+}(q) - w^{+}(p))$
4. $U(X_2) = \frac{w^{+}(p) + U(X_3)(w^{+}(q) - w^{+}(p))}{w^{+}(q)}$

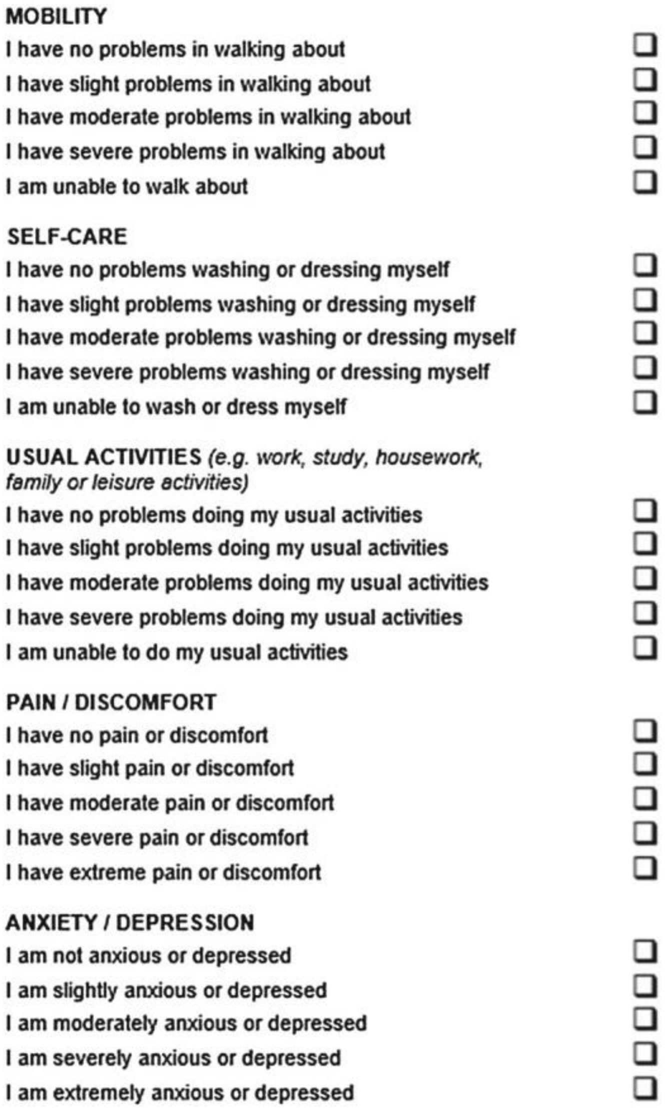
For reference point $'(X_1, T)'$, the following equation can be solved by applying formula 8, scaling $'U(X_1)'$ to 1 and by assuming multiplicativity in the following manner $((X_2, T), q; (X_3, T), 1 - q) = ((X_1, T), p; (X_3, T), 1 - p)$:

1. $U(X_1) * T^r \ - \ \lambda w^-(1-q)(U(X_1) * T^r - U(X_3) * T^r) - \lambda(1 - w^-(1-q))(U(X_1) * T^r - U(X_2) * T^r) = U(X_1) * T^r \ - \ \lambda w^-(1-p)(U(X_1) * T^r - U(X_3) * T^r)$
2. $w^-(1-q)(U(X_1) - U(X_3)) + (1 - w^-(1-q))(U(X_1) - U(X_2)) = w^-(1-p)(U(X_1) - U(X_3))$
3. $1 - w^-(1-q) * U(X_3) + w^-(1-q) * U(X_2) - U(X_2) = w^-(1-p) - \ w^-(1-p) * U(X_3)$
4. $(1 - w^-(1-q)) * U(X_2) = 1 - w^-(1-p) + \ w^-(1-p) * U(X_3) - \ w^-(1-q) * U(X_3)$
5. $U(X_2) = \frac{1-w^-(1-p)+U(X_3)(w^-(1-p)-w^-(1-q))}{1-w^-(1-q)}$

For reference point $'(X_2, T)'$, the following equation can be solved by applying formula 9, scaling $'U(X_1)'$ to 1 and by assuming multiplicativity in the following manner $((X_2, T), q; (X_3, T), 1 - q) = ((X_1, T), p; (X_3, T), 1 - p)$:

1. $U(X_2) * T^r - \lambda\, w^-(1-q)(U(X_2) * T^r - U(X_3) * T^r) = U(X_2) * T^r + w^+(p)(U(X_1) * T^r - U(X_2) * T^r) - \lambda\, w^-(1-p)(U(X_2) * T^r - U(X_3) * T^r)$
2. $-\lambda\, w^-(1-q) * U(X_2) + U(X_3) * \ \lambda\, w^-(1-q) = w^+(p) - U(X_2) * w^+(p) - \lambda\, w^-(1-p) * U(X_2) + \lambda\, w^-(1-p) * U(X_3)$
3. $U(X_2)\left(w^+(p) - \lambda\, w^-(1-q) + \lambda\, w^-(1-p)\right) = w^+(p) + \ U(X_3)(\lambda\, w^-(1-p) - \lambda\, w^-(1-q))$
4. $U(X_2) = \frac{w^+(p)+U(X_3)(\lambda w^-(1-p)-\lambda w^-(1-q))}{w^+(p)+ \lambda(w^-(1-p)-w^-(1-q))}$

Appendix 4: EQ-5D-5L system to describe health states on 5 dimensions and 5 levels (Devlin et al., 2017).



MOBILITY

I have no problems in walking about □
I have slight problems in walking about □
I have moderate problems in walking about □
I have severe problems in walking about □
I am unable to walk about □

SELF-CARE

I have no problems washing or dressing myself □
I have slight problems washing or dressing myself □
I have moderate problems washing or dressing myself □
I have severe problems washing or dressing myself □
I am unable to wash or dress myself □

USUAL ACTIVITIES *(e.g. work, study, housework, family or leisure activities)*

I have no problems doing my usual activities □
I have slight problems doing my usual activities □
I have moderate problems doing my usual activities □
I have severe problems doing my usual activities □
I am unable to do my usual activities □

PAIN / DISCOMFORT

I have no pain or discomfort □
I have slight pain or discomfort □
I have moderate pain or discomfort □
I have severe pain or discomfort □
I have extreme pain or discomfort □

ANXIETY / DEPRESSION

I am not anxious or depressed □
I am slightly anxious or depressed □
I am moderately anxious or depressed □
I am severely anxious or depressed □
I am extremely anxious or depressed □

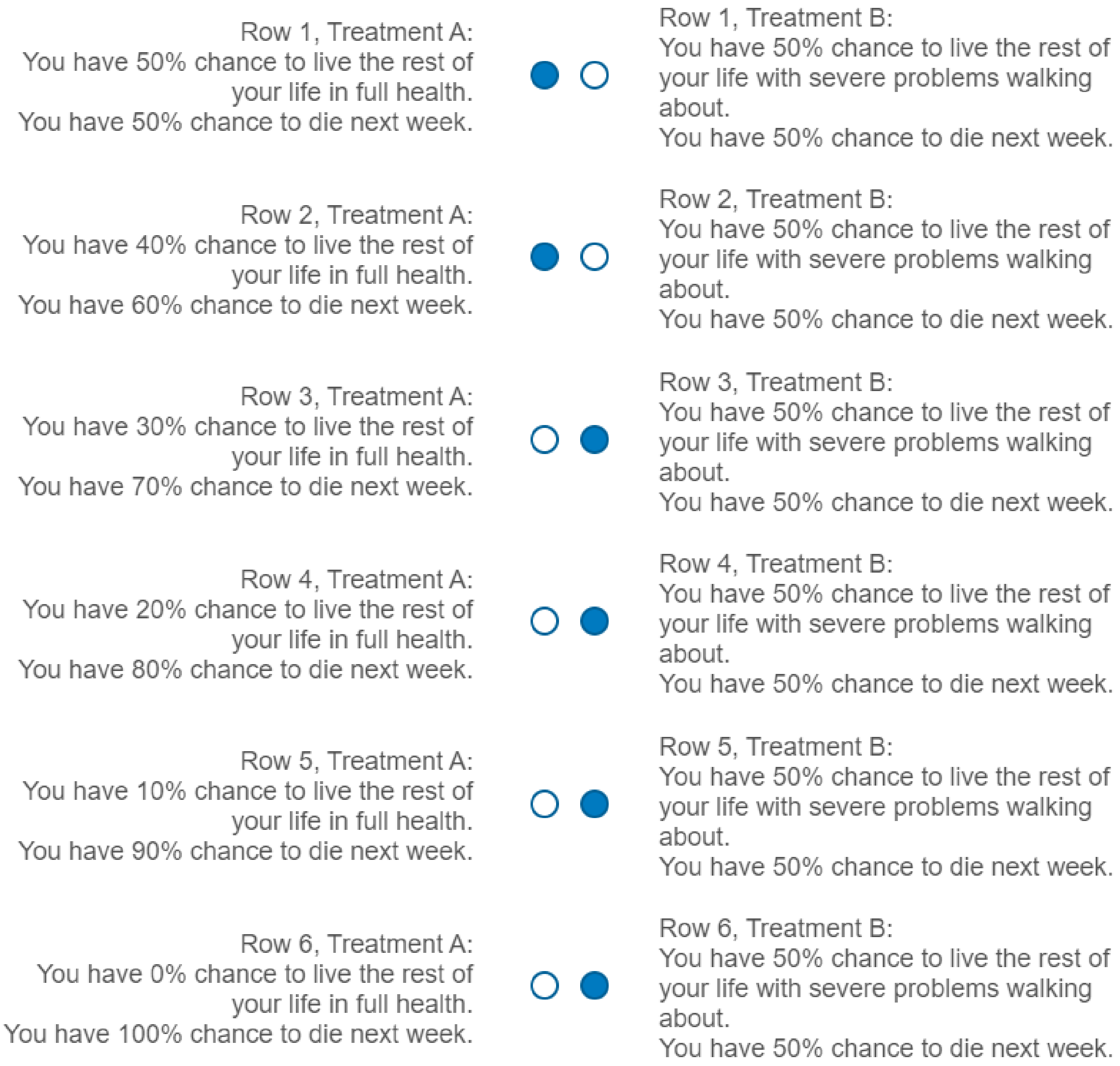
Appendix 5: Health state descriptions used in questionnaire:

Health state A ‘41111’
I have severe problems in walking about.

Health state B ‘41113’
I have severe problems in walking about and I am moderately anxious or depressed.

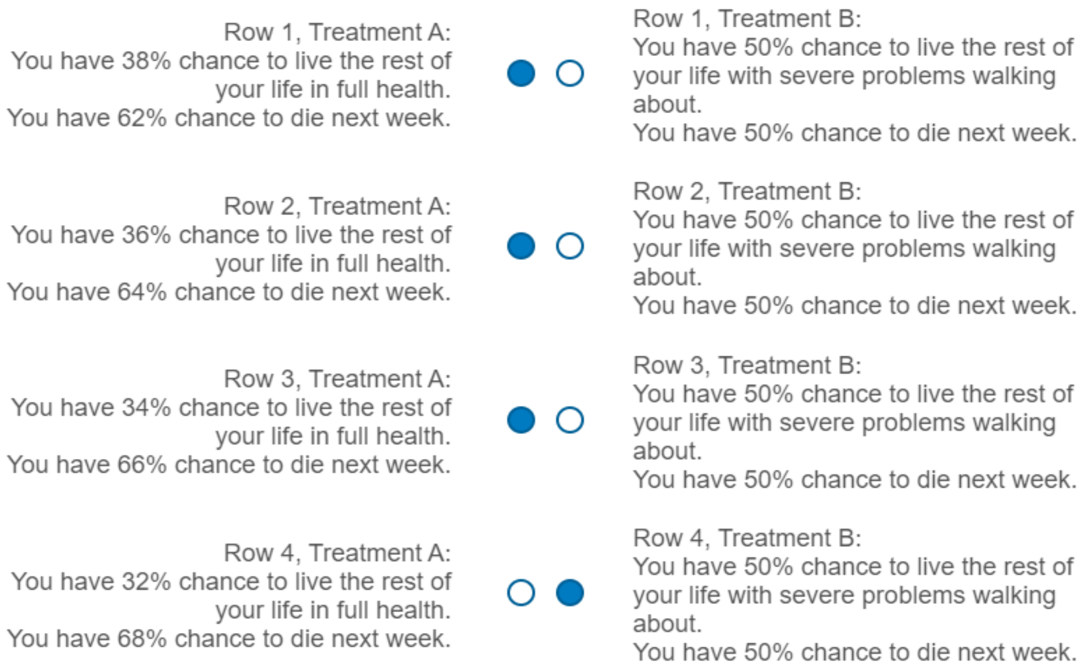
Appendix 6: Example choice list question used in questionnaire:

The first choice list (with increments for p of 10%) can correctly be filled out in the following manner by a hypothetical participant. His answer from the choice list implies that his indifference probability for gamble $(Full\ health, p; Death, 1-p)$ and gamble $(Health\ state\ A, 0.5; Death, 0.5)$ lies somewhere between p=0.3 and p=0.4:



Row 1, Treatment A: You have 50% chance to live the rest of your life in full health. You have 50% chance to die next week.	●	○	Row 1, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 2, Treatment A: You have 40% chance to live the rest of your life in full health. You have 60% chance to die next week.	●	○	Row 2, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 3, Treatment A: You have 30% chance to live the rest of your life in full health. You have 70% chance to die next week.	○	●	Row 3, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 4, Treatment A: You have 20% chance to live the rest of your life in full health. You have 80% chance to die next week.	○	●	Row 4, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 5, Treatment A: You have 10% chance to live the rest of your life in full health. You have 90% chance to die next week.	○	●	Row 5, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 6, Treatment A: You have 0% chance to live the rest of your life in full health. You have 100% chance to die next week.	○	●	Row 6, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.

Since the indifference probability of this individual lies somewhere between p=0.3 and p=0.4, a new choice list is presented where p varies from 0.32 till 0.38. If the participant fills out the choice list as outlined below, his indifference probability for both gambles is estimated at the mid-point of the probability of A for which he switches from treatment A to treatment B, which is equal to 0.33 in this scenario:

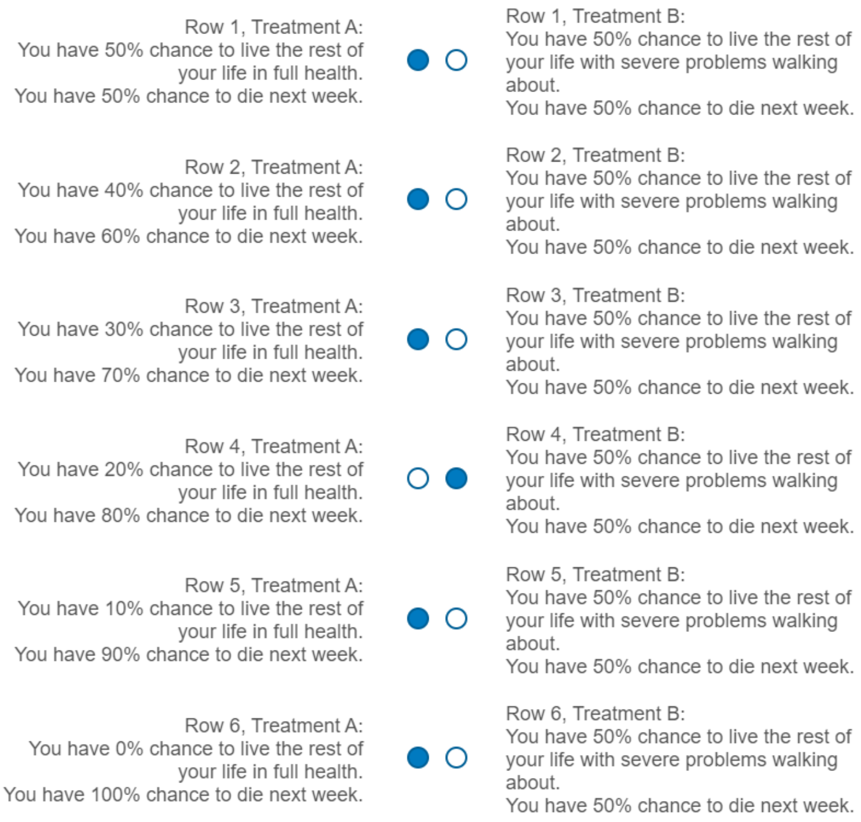


Row 1, Treatment A: You have 38% chance to live the rest of your life in full health. You have 62% chance to die next week.	● ○	Row 1, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 2, Treatment A: You have 36% chance to live the rest of your life in full health. You have 64% chance to die next week.	● ○	Row 2, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 3, Treatment A: You have 34% chance to live the rest of your life in full health. You have 66% chance to die next week.	● ○	Row 3, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 4, Treatment A: You have 32% chance to live the rest of your life in full health. You have 68% chance to die next week.	○ ●	Row 4, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.

Appendix 7:

https://qfreeaccountssjc1.az1.qualtrics.com/jfe/preview/previewId/95150548-33ce-4213-af66-c31900017588/SV_2lW1hMMt2Dur04C?Q_CHL=preview&Q_SurveyVersionID=current.

Appendix 8:



Row 1, Treatment A: You have 50% chance to live the rest of your life in full health. You have 50% chance to die next week.	● ○	Row 1, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 2, Treatment A: You have 40% chance to live the rest of your life in full health. You have 60% chance to die next week.	● ○	Row 2, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 3, Treatment A: You have 30% chance to live the rest of your life in full health. You have 70% chance to die next week.	● ○	Row 3, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 4, Treatment A: You have 20% chance to live the rest of your life in full health. You have 80% chance to die next week.	○ ●	Row 4, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 5, Treatment A: You have 10% chance to live the rest of your life in full health. You have 90% chance to die next week.	● ○	Row 5, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.
Row 6, Treatment A: You have 0% chance to live the rest of your life in full health. You have 100% chance to die next week.	● ○	Row 6, Treatment B: You have 50% chance to live the rest of your life with severe problems walking about. You have 50% chance to die next week.

Appendix 9:

Table 5: paired t-test chained and basic reference values health state A (n=57).	
Expected utility theory.	0.08 (0.026) ***
Prospect theory, reference point full health.	0.09 (0.024) ***
Prospect theory, reference point intermediate health state.	0.11 (0.022) ***
Prospect theory, reference point worst health state.	0.10 (0.023) ***

Significance at 10% level = *, significance at 5% level = **, significance at 1% level = ***.

Appendix 10:

Table 6: paired t-test chained and basic reference values health state A (n=45).	
Expected utility theory.	0.06 (0.028) **
Prospect theory, reference point full health.	0.06 (0.024) ***
Prospect theory, reference point intermediate health state.	0.08 (0.024) ***
Prospect theory, reference point worst health state.	0.07 (0.022) ***

Significance at 10% level = *, significance at 5% level = **, significance at 1% level = ***.

References

Abdellaoui, M. (2000). Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science*, *46*(11), 1497–1512. https://doi.org/10.1287/mnsc.46.11.1497.12080

Arrieta, A., García-Prado, A., González, P., & Pinto-Prades, J. L. (2017). Risk attitudes in medical decisions for others: An experimental approach. *Health Economics*, *26*, 97–113. https://doi.org/10.1002/hec.3628

Attema, A. E., Bleichrodt, H., L'Haridon, O., & Lipman, S. A. (2020). A comparison of individual and collective decision making for standard gamble and time trade-off. *European Journal of Health Economics*, *21*(3), 465–473. https://doi.org/10.1007/s10198-019-01155-x

Attema, A. E., & Brouwer, W. B. F. (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. https://doi.org/10.1016/j.joep.2013.07.009

Attema, A. E., Brouwer, W. B. F., & L'Haridon, O. (2013). Prospect theory in the health domain: A quantitative assessment. *Journal of Health Economics*, *32*(6), 1057–1065. https://doi.org/10.1016/j.jhealeco.2013.08.006

Barberis, N. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, *27*(1), 173–196. https://doi.org/10.1257/jep.27.1.173

Bartlett, J. G., Kotrlik, J. W., & Higgins, C. C. (2001). Organizational Research: Determining Appropriate Sample Size in Survey Research Appropriate Sample Size in Survey Research. *Information Technology, Learning, and Performance Journal*, *19*, 43–50.

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, *11*(5), 447–456. https://doi.org/10.1002/hec.688

Bleichrodt, H., Abellan-Perpiòan, J. M., Pinto-Prades, J. L., & Méndez-Martínez, I. (2007). Resolving Inconsistencies in Utility Measurement Under Risk: Tests of Generalizations of Expected Utility. *Management Science*, *53*(3), 469–482. https://doi.org/10.1287/mnsc.1060.0647

Bleichrodt, H., & Pinto, J. C. (2000). A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. *Management Science*, *46*(11), 1485–1496. https://doi.org/10.1287/mnsc.46.11.1485.12086

Bleichrodt, H., Pinto, J. C., & Wakker, P. P. (2001). Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility. *Management Science*, *47*(11), 1498–1514. https://doi.org/10.1287/mnsc.47.11.1498.10248

Bostic, R. W., Herrnstein, R., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior and Organization*, *13*(2), 193–212. https://doi.org/10.1016/0167-2681(90)90086-s

Brazier, J., Ara, R., Azzabi, I., Busschbach, J. J. V., Chevrou-Severac, H., Crawford, B. J., Cruz, L. N., Karnon, J., Lloyd, A. R., Paisley, S., & Pickard, A. S. (2019). Identification, Review, and Use of Health State Utilities in Cost-Effectiveness Models: An ISPOR Good Practices for Outcomes Research Task Force Report. *Value in Health*, *22*(3), 267–275. https://doi.org/10.1016/j.jval.2019.01.004

Brazier, J., & Ratcliffe, J. (2017). Measurement and Valuation of Health for Economic Evaluation. In S. R. Quah (Ed.), *International Encyclopedia of Public Health* (pp. 586-593). https://doi.org/10.1016/B978-0-12-803678-5.00457-4

Brown, A., Imai, T., Vieider, F. M., & Camerer, C. F. (2021). *Meta-Analysis of Empirical Estimates of Loss-Aversion* (CESifo Working Paper No. 8848). Retrieved from website of Social Science Research Network: https://doi.org/10.2139/ssrn.3772089

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

D'Agostino, R. B., & Belanger, A. J. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician*, *44*(4), 316–321. https://doi.org/10.1080/00031305.1990.10475751

Devlin, N., Shah, K., Feng, Y., Mulhern, B., & Van Hout, B. (2017). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics*, *27*(1), 7–22. https://doi.org/10.1002/hec.3564

Feeny, D., & Eng, K. (2005). A test of prospect theory. *International Journal of Technology Assessment in Health Care*. *21*(4), 511–516. https://doi.org/10.1017/s0266462305050713

Gonzalez, R., & Wu, G. Y. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, *38*(1), 129–166. https://doi.org/10.1006/cogp.1998.0710

Jones-Lee, M., Loomes, G., & Philips, P. R. (1995). Valuing the prevention of non-fatal road injuries: contingent valuation vs. standard gambles. *Oxford Economic Papers*, *47*(4), 676–695. https://doi.org/10.1093/oxfordjournals.oep.a042193

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263. https://doi.org/10.2307/1914185

Kim, D. H., & Basu, A. (2021). How Does Cost-Effectiveness Analysis Inform Health Care Decisions? *AMA Journal of Ethics*, *23*(8), E639-647. https://doi.org/10.1001/amajethics.2021.639

Law, A. V., Pathak, D. S., & McCord, M. R. (1998). Health status utility assessment by standard gamble: a comparison of the probability equivalence and the lottery equivalence approaches. *Pharmaceutical Research*, *15*(1), 105–109. https://doi.org/10.1023/a:1011913123135

McCord, M. A., & De Neufville, R. (1986). "Lottery Equivalents": Reduction of the Certainty Effect Problem in Utility Assessment. *Management Science*, *32*(1), 56–60. https://doi.org/10.1287/mnsc.32.1.56

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*(1), 3–20. https://doi.org/10.1037/0096-3445.117.1.3

Oliver, A. (2003). The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Journal of Health Economics*, *22*(4), 659–674. https://doi.org/10.1016/s0167-6296(03)00023-7

Oliver, A. (2004). Testing the internal consistency of the standard gamble in 'success' and 'failure' frames. *Social Science & Medicine*, *58*(11), 2219–2229. https://doi.org/10.1016/j.socscimed.2003.08.024

Oliver, A. (2005). Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Economics*, *14*(2), 149–159. https://doi.org/10.1002/hec.889

Ott, R., & Longnecker, M. (2015). *An Introduction to Statistical Methods and Data Analysis* (7th ed.). Boston, MA: Cengage Learning.

Rodríguez-Míguez, E., Pinto-Prades, J. L., & Mosquera-Nogueira, J. (2019). Eliciting Health State Utilities Using Paired-Gamble Methods: The Role of the Starting Point. *Value in Health*, *22*(4), 446–452. https://doi.org/10.1016/j.jval.2019.01.007

Ross, S. M. (2020). *Introduction to Probability and Statistics for Engineers and Scientists* (5th ed.). London, England: Elsevier.

Salomon, J. A. (2014). Valuing Health States, Techniques for. In A. J. Culyer (Ed.), *Encyclopedia of Health Economics* (pp. 454–458). https://doi.org/10.1016/b978-0-12-375678-7.00502-2

Schmidt, U., & Zank, H. (2012). A genuine foundation for prospect theory. *Journal of Risk and Uncertainty*, *45*(2), 97–113. https://doi.org/10.1007/s11166-012-9150-8

Strack, F. (1992). “Order Effects” in Survey Research: Activation and Information Functions of Preceding Questions. In N. Schwarz & S. Sudman (Eds.), *Context Effects in Social and Psychological Research* (pp. 23–34). https://doi.org/10.1007/978-1-4612-2848-6_3

Stratton, S. J. (2021). Population Research: Convenience sampling strategies. *Prehospital and Disaster Medicine*, *36*(4), 373–374. https://doi.org/10.1017/s1049023x21000649

Swedberg, R. (2020). Exploratory Research. In C. Elman, J. Gerring & J. Mahoney (Eds.), *The Production of Knowledge: Enhancing progress in social science* (pp. 17–41). https://doi.org/10.1017/9781108762519.002

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/bf00122574

Van Osch, S. M. C., 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*(4), 338–346. https://doi.org/10.1177/0272989x06290484

Varian, H. R. (2010). *Intermediate microeconomics: A modern approach.* New York, NY: W. W. Norton & Company.

Versteegh, M., Vermeulen, K. M., Evers, S. M. a. A., De Wit, G. A., Prenger, R., & Stolk, E. A. (2016). Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health*, *19*(4), 343–352. https://doi.org/10.1016/j.jval.2016.01.003